# Chapter one

This thesis was presented in a journal style. The meta-analysis presented in chapter two has been reproduced with permission from Springer Nature after publication in Education Psychology Review, chapter three is currently being resubmitted, and chapter four was written in journal article form to maintain a similar style across chapters.

Many recent suggested best practice in education have focused on the long-term retention of key mathematical knowledge ([Ofsted, 2021](#ref-ofsted2021)). 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](#ref-karpicke2012)). 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](#ref-hopkins2016)). Already, spaced retrieval practice is in regular use by many students in the United Kingdom. Online Platforms such as ([ARC Education, n.d.](#ref-arceducation)) and ([*Sparx Maths*, n.d.](#ref-sparxma)) claim to harness the spacing and testing effect to improve retention for pupils. In particular, *Sparx Maths* ([n.d.](#ref-sparxma)) states that they “*ensure the practice uses spaced repetition and interleaving to support a change in students’ long-term memories*” and their software is “*is proven by The University of Cambridge to significantly boost grades*”. 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](#ref-brown2021)). 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](#ref-gidaropoulos2021)). With sites such as Sparx Maths being in use by more than 2 million students in more than 2,200 schools ([*Sparx Maths*, n.d.](#ref-sparxma)), 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](#ref-ofsted2021)). 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.

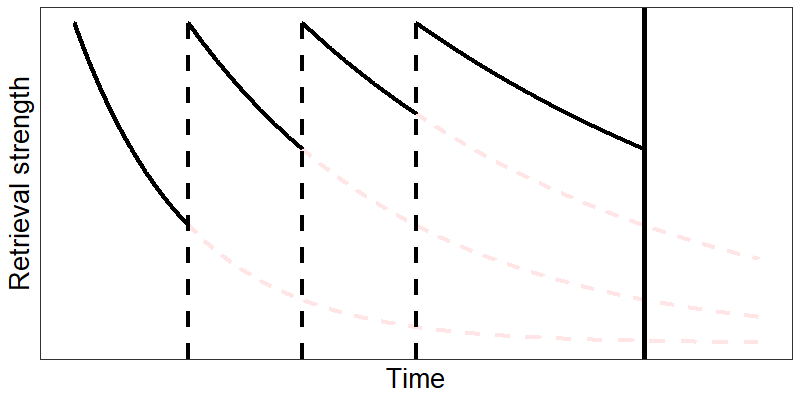
# Spacing effect and distributed practice

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](#ref-delaney2010)). The lag effect is the change in retention when two different spacing routines are compared ([Küpper-Tetzel, 2014](#ref-küpper-tetzel2014)). Distributed practice is often used as an umbrella term for interventions that utilise the spacing or lag effects ([Benjamin & Tullis, 2010](#ref-benjamin2010)). A meta-meta-analysis by Hattie ([2008](#ref-hattie2008visible)) into different types of educational interventions found spaced versus massed practice to be one of the most effective. They found an 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 ([1964](#ref-ebbinghaus1964)) 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](#fig-fogcur)); these results have been replicated successfully ([Murre & Dros, 2015](#ref-murre2015)). The spacing effect has been the subject of hundreds of experiments ([Cepeda et al., 2006](#ref-cepeda2006); [Hattie, 2008](#ref-hattie2008visible)), proving to be a robust effect found across age groups, tasks, and species ([Walsh et al., 2018](#ref-walsh2018)). However, most prior experiments on the spacing effect have investigated learning verbal material ([Delaney et al., 2010](#ref-delaney2010), provides a critical review).

Figure 1

Forgetting curve showing the change in retrieval strength as a function of time

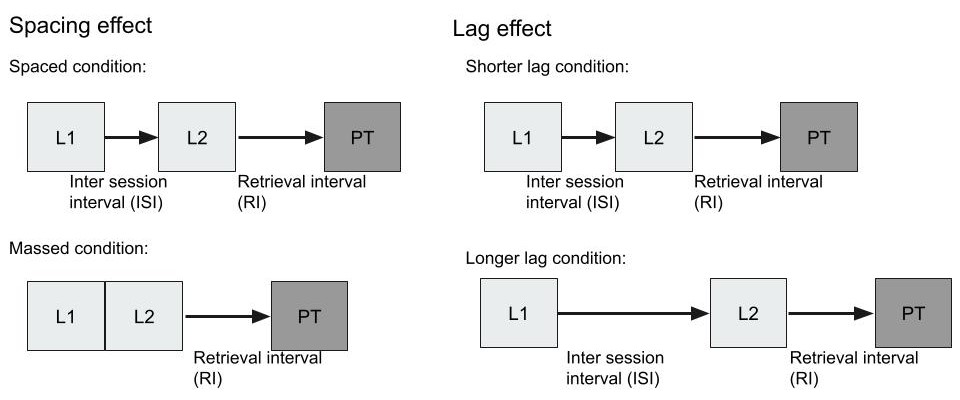


*Note*. Retrieval strength signifies how easily the item can be retrieved from long-term memory. The thick vertical line represents a post-test. The dashed vertical lines signify retrieval events, the lighter dashed lines show how the memory would have continued to decay without the retrieval events.

Spacing effect experiments can be conducted either within a single session or across multiple sessions ([Küpper-Tetzel, 2014](#ref-küpper-tetzel2014)). Within-session experiments look at the temporal spacing of stimuli within a single session, whilst across 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](#ref-foster2019); [Rea & Modigliani, 1985](#ref-rea1985)), while this is much more common within other domains such as verbal learning ([Delaney et al., 2010](#ref-delaney2010)). Therefore, this review will focus on across-session spacing experiments. In the most basic across-session spacing effect experiment (see [Figure 2](#fig-spacing)), 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 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](#ref-cepeda2008)). 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 labeled the Glenberg surface ([Delaney et al., 2010](#ref-delaney2010)). 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](#ref-rea1985)). 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](#ref-bjork2011)). However, a recent meta-analysis looking at spaced retrieval practice found no significant difference between expanding and uniform designs ([Latimier et al., 2021](#ref-latimier2021)). 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](#ref-emeny2021)). In contrast, students who followed the massed schedule tended to be overconfident with their predictions. Emeny et al. ([2021](#ref-emeny2021)) 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.

## 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](#ref-delaney2010); [Dempster, 1988](#ref-dempster1988); [Walsh et al., 2018](#ref-walsh2018)). 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](#ref-delaney2010)) for a critical review of within-session verbal learning spacing experiments and Küpper-Tetzel ([2014](#ref-küpper-tetzel2014)) 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](#ref-thios1976)). The degree to which study-phase retrieval improves retention is dependent on the difficulty of recall ([Küpper-Tetzel, 2014](#ref-küpper-tetzel2014)). The more difficult an 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](#ref-bjork2011)). Evidence for this phenomenon was found by Magliero ([1983](#ref-magliero1983)) 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](#ref-hintzman1973)). Further support was found by Wahlheim et al. ([2014](#ref-wahlheim2014)), 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](#ref-glenberg1979)). 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](#ref-küpper-tetzel2014)). Küpper-Tetzel ([2014](#ref-küpper-tetzel2014)) 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](#ref-dempster1988)). An early, though notable, failure to empirically support contextual variability was performed by Ross and Landauer ([1978](#ref-ross1978)). 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](#ref-lohnas2011)) 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 *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 ([Hintzman, 1974](#ref-hintzman1974)). 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](#ref-shaughnessy1972)). 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](#ref-koval2019)). 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](#ref-delaney2010)). 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](#ref-benjamin2010)). 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](#ref-sweller1988)), *Working Memory Resource Depletion* has been offered as an alternative theory to explain some forms of the spacing effect Chen & Kalyuga ([2019](#ref-chen2019)). They suggest that during massed practice working memory resources are depleted which leads to reduced performance. While in the spaced condition, participants have the opportunity to 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 in order to complete a task ([Chen et al., 2023](#ref-chen2023)). 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](#ref-chen2024)), 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](#ref-delaney2010)). 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](#ref-chen2024))’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 leveling off. Similarly, the rest periods in Chen et al. ([2024](#ref-chen2024)) 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](#ref-walsh2018)). 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](#ref-hintzman1973)). The second model, first introduced by Pavlik and Anderson ([2005](#ref-pavlik2005)), 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](#ref-raaijmakers2003)), 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. Importantly, only the working memory depletion hypothesis takes into account the

# 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 breifly 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](#ref-roediger2006)). 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](#ref-yang2021)). 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.

# Theories of the testing effect

Many theories aim to explain the testing effect ([Rowland, 2014](#ref-rowland2014)), 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](#ref-pyc2009)). The mechanism in which increased effort leads to increased retrieval is expanded upon by the dual memory theory ([Rickard & Pan, 2018](#ref-rickard2018)). 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 modeled 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](#ref-carpenter2009)). In one experiment, participants learnt word pairs. These pairings were either strongly associated (Toast - Bread) or weakly associated (Basket - Bread) ([Carpenter, 2009](#ref-carpenter2009)). 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](#ref-blaxton1989); [Morris et al., 1977](#ref-morris1977levels)). However, an experiment designed to evaluate transfer-appropriate processing against the elaborative retrieval hypothesis found little evidence for transfer-appropriate processing ([Carpenter & DeLosh, 2006](#ref-carpenter2006impoverished)). 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](#ref-karpicke2014)) 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](#ref-whiffen2017)). 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.

# 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](#ref-latimier2021)). 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.

## Typical paradigms in mathematics spaced (retrieval) 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](#ref-küpper-tetzel2014)). 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](#ref-rohrer2007); [Rohrer & Taylor, 2006](#ref-rohrer2006)), 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](#ref-bego2017); [Crissinger, 2015](#ref-crissinger2015); [Lyle et al., 2020](#ref-lyle2020)). 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](#ref-chen2021)).

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](#ref-rohrer2006)). 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](#ref-rohrer2007)).

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](#ref-hopkins2016)) 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](#ref-hopkins2016); [Lyle et al., 2020](#ref-lyle2020)). 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](#ref-beagley2020)).

## 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 ([Nazari & Ebersbach, 2019](#ref-nazari2019)). 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](#ref-donovan1999)). They found that overall complexity of the material was correlated with lower effect sizes.

# 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](#ref-donovan1999)) 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](#ref-chen2024)) 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.

## Defining complexity

In the mathematics learning domain task complexity is often defined procedurally or conceptually ([Crooks & Alibali, 2014](#ref-crooks2014)). 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](#ref-crooks2014)). 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](#ref-roediger2012)). Additionally, a procedure can be taught in isolation, while conceptual knowledge, which requires links between topics by definition, cannot ([Hiebert & Lefevre, 1986](#ref-hiebert1986procedural)). 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](#ref-crooks2014)). 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](#ref-hiebert1986procedural)). 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](#ref-rohrer2007)). 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 four 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.

## 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](#ref-cepeda2006)), 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](#ref-bjork2011)). 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.

## 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](#ref-donovan1999)). 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 adressed through the meta-analysis in chapter two. 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 three. Does element interactivity affect the efficacy of spaced practice? This question will be adressed in chapter four.

# References

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

Beagley, J., & 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>

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>

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>

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>

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*, 268276.

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., Castro-Alonso, J. C., Paas, F., Sweller, J., & John;. (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*, 102281. <https://doi.org/10.1016/j.cedpsych.2024.102281>

Chen, O., & Kalyuga, S. (2019). *Cognitive load theory, spacing effect, and working memory resources depletion* (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., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educ. Psychol. Rev.*, *35*(2), 63. <https://doi.org/10.1007/s10648-023-09782-w>

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., 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*.

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*.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* Dover.

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

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>

Gidaropoulos, A. (2021). *Hegartymaths: 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>

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

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., 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>

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. *Educ. Psychol. Rev.*, *28*(4), 853–873. <https://doi.org/10.1007/s10648-015-9349-8>

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., 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>

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>

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

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., 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. *Educ. Psychol. Rev.*, *32*(1), 277–295. <https://doi.org/10.1007/s10648-019-09489-x>

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>

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), 519533.

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>

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

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

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

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. *Cogn. Sci.*, *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>

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

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. *J. Appl. Res. Mem. Cogn.*, *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. *Appl. Cogn. Psychol.*, *20*(9), 1209–1224. <https://doi.org/10.1002/acp.1266>

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>

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>

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

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>

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>

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. *J. Exp. Psychol. Learn. Mem. Cogn.*, *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. *Cogn. Sci.*, *42 Suppl 3*, 644–691. <https://doi.org/10.1111/cogs.12602>

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>

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>