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

Ewan Murray[1], Aidan J. Horner[1,2], and Silke M. Göbel[1,3]

[1]Department of Psychology, University of York, UK

[2]York Biomedical Research Institute, University of York, UK

[3]CREATE & Department of Special Needs Education, University of Oslo

Author Note

Correspondence concerning this article should be addressed to Ewan Murray, Department of Psychology, University of York, UK, Email: ewan.murray@york.ac.uk

Abstract

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

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

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

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

The Spacing Effect

The spacing effect is a robust phenomenon in learning and memory research. Previous meta-analyses have found large beneficial effects to learning outcomes across hundreds of studies (Cepeda et al., 2006; Donovan & Radosevich, 1999; Latimier et al., 2021), and it has been subject to multiple reviews (Delaney et al., 2010; Küpper-Tetzel, 2014; Maddox, 2016). In the domain of mathematics, there have been mixed results. Significant benefits have been reported for learning arithmetic facts (Schutte et al., 2015), algebra (Gay, 1973), geometry (Yazdani & Zebrowski, 2006), simple procedures (Rohrer & Taylor, 2006), and more complex bodies of knowledge such as calculus (Hopkins et al., 2016; Lyle et al., 2020a, 2022). Significant positive effects have been found across age groups including primary school (Chen & Kalyuga, 2019), secondary school (Emeny et al., 2021; Nazari & Ebersbach, 2019) and higher education (Hopkins et al., 2016; Lyle et al., 2020a, 2022).

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

In summary, there is evidence that spaced practice can be beneficial for mathematics learning, however, there may be additional factors that limit the effects of spacing or require the practice to be implemented differently. We will focus on one such factor in the following experiments, the complexity of the material.

Complexity and spacing

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

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

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

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

We had three hypotheses. Firstly, spaced retrieval would lead to a greater retention than massed retrieval across all conditions. We did not expect the spacing effect to fully disappear or negatively affect learning. This hypothesis predicts a significant main effect of spacing. Secondly, participants would have lower retention of higher complexity material. We predicted that as there were more connected steps to recall and use, participants in the higher complexity condition would have lower performance on the post-test. This hypothesis predicts a significant main effect of complexity. Thirdly, there would be an interaction between spacing and complexity. Specifically, there would be a significantly smaller effect of spacing in the high complexity condition than in the low complexity condition.

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

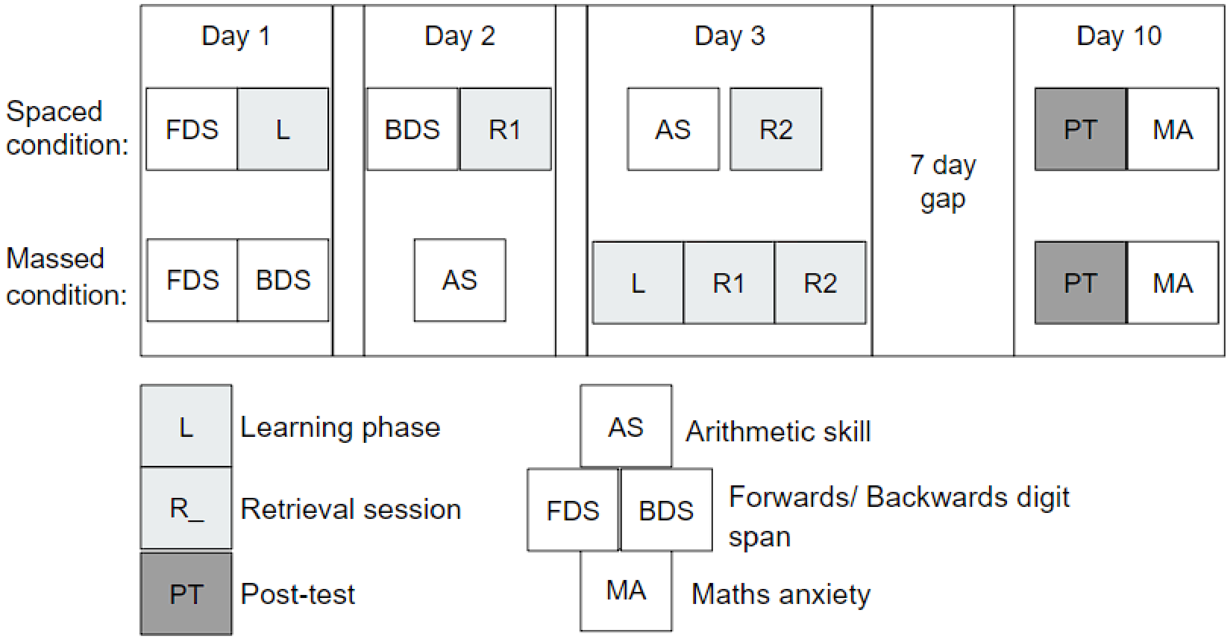
Experiment one

Method

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

Figure 1

Experimental Design



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

Participants

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

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

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

Material

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

In both the learning session, and the retrieval sessions, participants were cued with the instructions “Apply the n-step procedure to the numbers a and b” with n being either “two” or “three” and *a* and *b* two integers such that the result of applying the procedure is a whole number. In the learning session, the steps were displayed while participants applied the procedure, while in the retrieval sessions the participants had to retrieve the steps to apply the procedure. In all three sessions, when an incorrect answer was submitted participants were given feedback in the form of the correct answer and were shown the steps of the procedure again. Each session consisted of nine trials for the two-step and nine trials for the three-step procedure. Retrieval accuracy was defined as the mean accuracy across the two retrieval sessions.

Table 1

Procedures learnt by participants

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

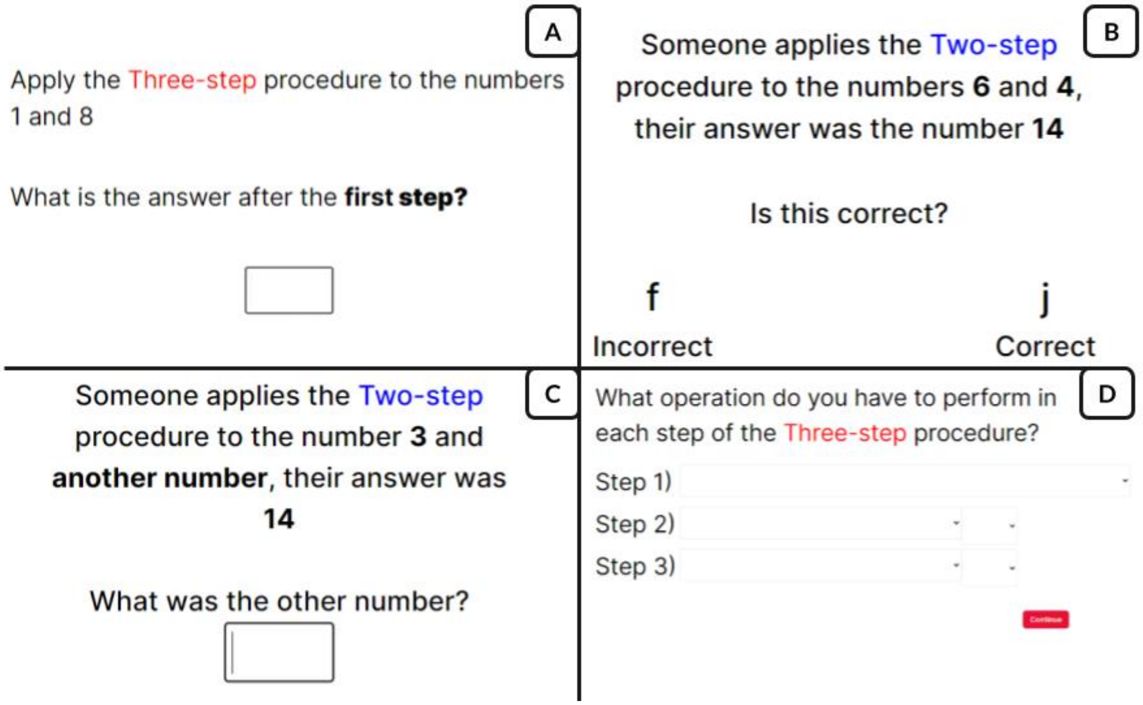
Note. Table displaying the procedures used in the experiments. * = used only in experiment one, ** = used only in experiment two

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

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

Figure 2

Post-test sub-tasks



Apply the Three-step procedure to the numbers 1 and 8

What is the answer after the **first step?**

Someone applies the Two-step procedure to the numbers **6** and **4**, their answer was the number **14**

Is this correct?

Someone applies the Two-step procedure to the number **3** and **another number**, their answer was **14**

What was the other number?

What operation do you have to perform in each step of the Three-step procedure?

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

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

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

Mathematics anxiety was measured using an online version of the Mathematics Anxiety Scale–UK (MAS-UK) (Hunt et al., 2011). This scale was specifically designed for the British undergraduate student population, which matched our target population. It consisted of twenty-three hypothetical scenarios designed to measure participants' mathematics anxiety. For each scenario, participants rated their predicted anxiety on a on a five-point scale (with the answer options: 'not at all', 'slightly', 'a fair amount', 'much', 'very much'), where a higher score meant they felt more anxious at the prospect of doing that activity. For example, participants were asked to imagine how anxious they would feel when a situation occurs such as "Having someone watch you multiply 12 x 23 on paper". The final mathematics anxiety score was the sum of their answers.

Procedure

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

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

In another study involving spacing, there was more attrition in the spaced condition, than the massed condition, which meant the original analyses could not be run (Nazari & Ebersbach, 2018). To reduce the possibility of this for the present study we ensured that participants in the massed and spaced condition both had to complete four sessions over ten days. The digit span and arithmetic fluency tasks were used to pad the remaining sessions to ensure that participants had to return to the experiment an equal number of times across both conditions.

Analysis

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

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

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

Results

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

Table 2

Experiment one descriptive statistics

Spacing	Procedure	Overall Score	Retrieval Accuracy	Working Memory	Mathematics Anxiety	Arithmetic Fluency
Massed	Two-step	0.561 (0.276)	0.843 (0.109)	19.971 (3.362)	51.118 (10.117)	69.265 (23.839)
	Three-step	0.608 (0.318)	0.863 (0.120)			
Spaced	Two-step	0.698 (0.229)	0.813 (0.079)	18.972 (3.468)	52.361 (14.333)	70.389 (23.626)
	Three-step	0.734 (0.267)	0.844 (0.086)			

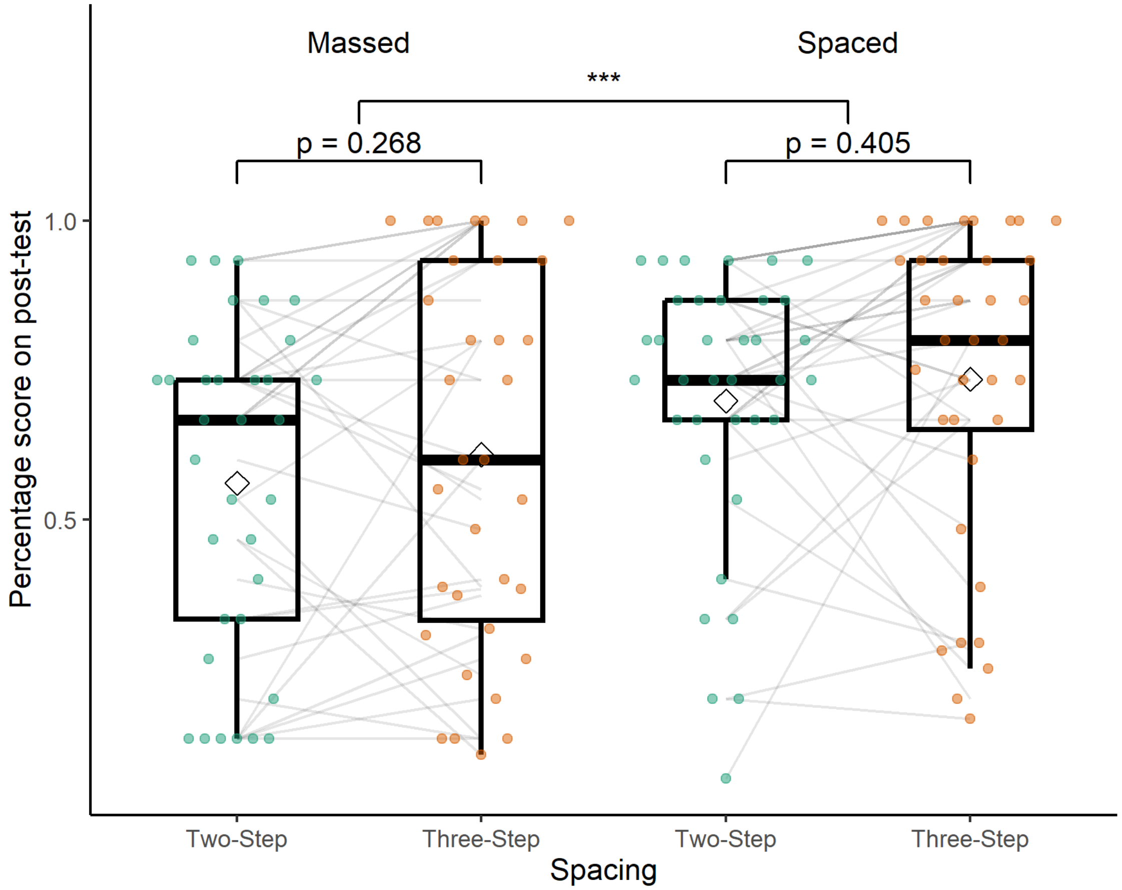
Note. Table displaying the mean and standard deviation for the percentage score on the post-test, Retrieval Accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, mathematics anxiety and arithmetic fluency tasks

A two (spaced/massed) by two (two-step/ three-step procedure) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable. There was a significant main effect of spacing, *F*(1,68) = 5.135, p = 0.027, η^2 = 0.056 (BF_{10} = 2.354). The percentage overall score on the post-test was significantly higher in the spaced (*M* = 0.716, *SD* = 0.248) than in the massed condition (M = 0.585, *SD* = 0.297). There was no significant main effect of complexity, *F*(1,68) = 1.926, *p* = 0.17, η^2 = 0.006 (BF_{10} = 0.421). There was no evidence of an interaction between spacing and complexity, *F*(1,68) = 0.035, *p* = 0.853, η^2 = 0 (BF_{10} = 0.248). We therefore saw a clear effect of spacing, however the lack of difference between the two-step and three-step procedure suggests that we may not have systematically varied complexity sufficiently.

Finally, we assessed whether the spacing effect was present in both the complexity conditions separately. Post-hoc t-tests revealed a significant difference between massed (*M* = 0.561, *SD* = 0.276) and spaced (*M* = 0.698, *SD* = 0.229) performance on the post-test for the two-step procedures (*t*(34) = -2.258, *p* = 0.027, BF_{10} = 2.146) after Bonferroni adjustment. However, there was no significant difference between massed and spaced three-step procedures (*t*(34) = -1.79, *p* = 0.078, BF_{10} = 0.969), after Bonferroni adjustment.

Figure 3

Experiment one - Main effect of Spacing



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

Table 3

Experiment one break down of post-test subtasks

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

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

Exploratory analyses

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

Table 4

Experiment one exploratory variables t-test

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

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

Discussion

In this experiment we tested how changes to the number of steps in a procedure affected the efficacy of spaced retrieval practice. Our first hypothesis was that participants would have lower retention of high complexity material. However, the main effect of complexity was not significant and, numerically, participants appeared to perform better in the higher complexity condition. Secondly, we predicted that across all levels of complexity spaced retrieval would lead to a greater retention than massed retrieval. This was not the case. There was an overall significant effect of spacing, although, post-hoc results only showed significant evidence for a spacing effect in the lower complexity condition and the evidence was not significant in the higher complexity condition. The Bayes factor for the comparison between spaced and massed three-step procedures was close to one. This evidence suggests that it is equally likely that there is a difference as that there is not a difference. Finally, based on previous evidence we predicted there would be a larger spacing effect for lower complexity material than higher complexity material. We found no evidence for an interaction between spacing and complexity. In the next experiment we aimed to increase the difference in complexity by adding additional steps to the higher complexity procedure.

Experiment two

Method

Participants

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

Materials

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

Procedure

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

Results

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

Table 5

Experiment two descriptive statistics

Spacing	Procedure	Retrieval Accuracy	Overall Score	Working Memory	Mathematics Anxiety	Arithmetic Fluency
Massed	Two-step	0.949 (0.106)	0.659 (0.225)	19.088 (3.297)	54.412 (12.225)	62.853 (18.868)
	Five-step	0.959 (0.126)	0.495 (0.251)			
Spaced	Two-step	0.969 (0.102)	0.852 (0.108)	19.824 (3.407)	49.029 (14.379)	75.794 (24.733)
	Five-step	0.928 (0.095)	0.7 (0.228)			

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

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

Table 6

Experiment two Robust ANOVA

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

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

There was a large and significant main effect of spacing, $F(1,63) = 22.523$, $p < 0.001$, $\eta^2 = 0.184$ ($BF_{10} = 750.646$). The percentage overall score on the post-test was significantly higher in the spaced condition ($M = 0.776$, $SD = 0.193$) than in the massed condition ($M = 0.577$, $SD = 0.251$). There was a significant medium to large main effect of complexity, $F(1,63) = 24.308$, $p < 0.001$, $\eta^2 = 0.124$ ($BF_{10} = 4330.719$), such that participants performed better on the two-step procedures than the five-step. There was no evidence of an interaction between spacing and complexity, $F(1,63) = 0.036$, $p = 0.85$, $\eta^2 = 0$ ($BF_{10} = 0.265$).

Table 7

Experiment two break down of post-test subtasks

Spacing	Procedure	Basic	Correct (y/n)	Recognise Steps	Reverse Procedure
Massed	Two-step	0.525 (0.266)	0.755 (0.263)	0.814 (0.309)	0.676 (0.676)
	Five-step	0.363 (0.291)	0.647 (0.306)	0.732 (0.311)	0.373 (0.373)
Spaced	Two-step	0.681 (0.190)	0.922 (0.165)	0.902 (0.225)	0.863 (0.863)
	Five-step	0.578 (0.291)	0.804 (0.219)	0.904 (0.246)	0.637 (0.637)

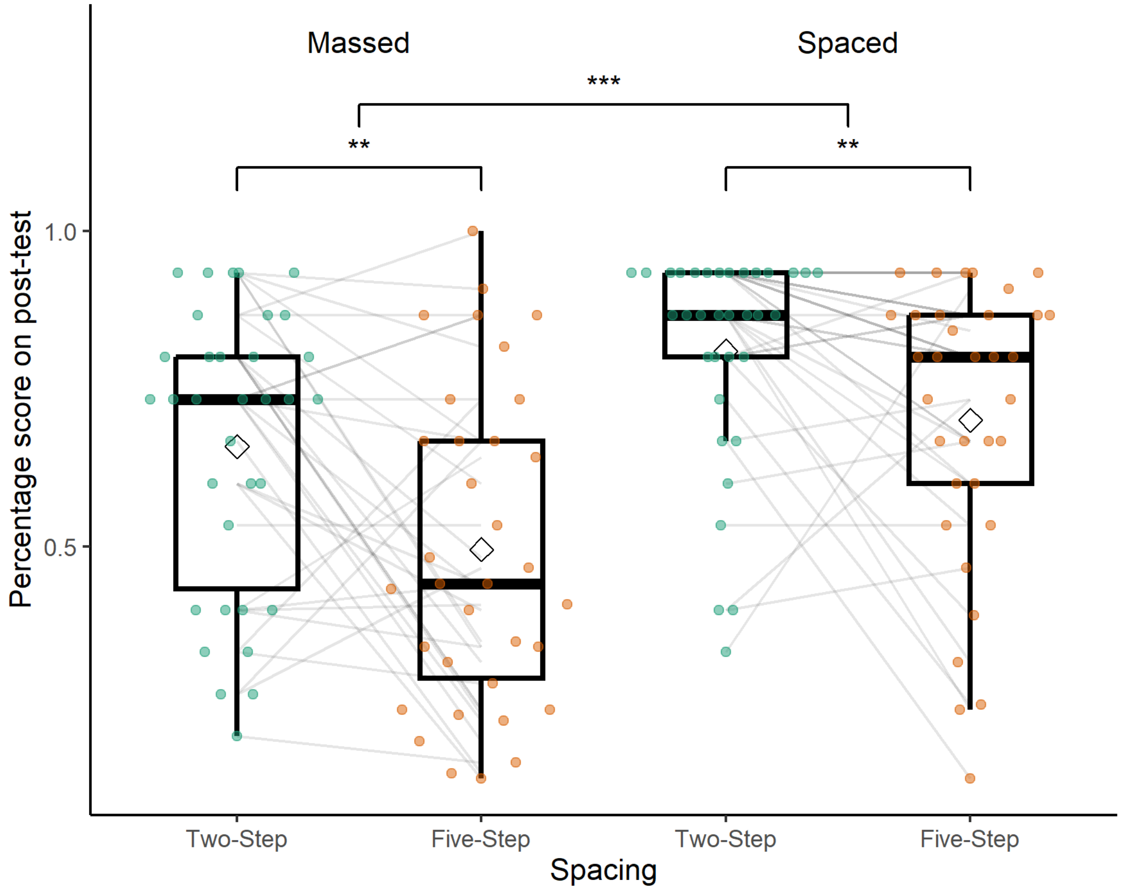
Note. Table displaying the mean and standard deviation for the percentage score on the post-test, broken down by the sub-tasks

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

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

Figure 4

Experiment two - Main effect of Spacing



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

Exploratory analyses

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

Table 8

Experiment two exploratory variables t-tests

	Statistic	df	p	p.adj
Arithmetic Fluency	-2.426	61.692	0.018	0.073
Mathematics Anxiety	1.663	64.336	0.101	0.404
Retrieval Accuracy	-0.778	65.868	0.440	1.000
Working Memory	-0.904	65.930	0.369	1.000

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

Discussion

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

General discussion

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

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

Although we found evidence for a main effect of spacing across both experiments, the post-hoc tests, performed for completeness, suggested the three-step procedure (experiment one) did not show a significant spacing effect (and the Bayes Factor was inconclusive). Although this post-hoc test could simply be a false negative, the deficient processing account (Hintzman, 1974) could explain the lack of spacing effect for the three-step procedure. If the three-step procedure required additional processing it may have not been negatively affected by the deficient processing in massed practice. The two-step procedure may have been simple enough that processing dropped over a small number of trials, therefore those participants benefited from the boost to attention provided by spaced practice. However, if that were the case then we would have expected a similar effect for the five-step procedure. One possibility is that the five-step procedure may have had too many elements to be processed at once, therefore the lack of deficient processing in the massed condition may not have had an effect. Overall, there is little evidence that this is the case, a false negative is the simpler explanation, particularly given the inconclusive Bayes Factor, the presence of a spacing effect in the other three conditions, and a main effect of spacing in experiment one.

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

Our third hypothesis was that there would be an interaction between spacing and complexity. More specifically, that there would be a smaller effect of spacing in the high complexity condition than in the low complexity condition. We had two reasons to predict this. Firstly, considering the study-phase retrieval hypothesis, longer procedures may be more difficult to retrieve. More difficult retrievals could in turn lead to fewer or more inaccurate retrievals of the original study-phase memory trace, negating any benefit of spacing. Secondly, considering the deficient processing account, we hypothesized that adding additional steps to a procedure may increase the level of processing required to retrieve the procedure. Given a finite level of maximum potential processing, this may mean there is no beneficial effect to memory of additional processing in the spaced condition. We found no evidence for an interaction between procedural complexity and the spacing effect in either experiment. Furthermore, the results of the Bayesian ANOVA suggest that there was strong evidence for the null hypothesis, that there was no interaction between spacing and procedural complexity. This provides evidence against our initial hypothesis and suggests that spacing can be effective for varying levels of procedural complexity: when operationalised as the number of steps in a procedure. While this goes against our initial hypothesis it suggests that the spacing effect is robust to variations in procedural complexity and supports its use in the teaching and learning of mathematics.

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

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

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

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

Conclusion

Across two experiments, we found significant main effects of spacing. Critically, the spacing effect was not modulated by procedural complexity. Our results suggest that, when defined as the number of steps in a procedure, procedural complexity does not affect the efficacy of spaced retrieval practice for mathematics. This provides further support to the robust nature of the spacing effect that makes it a valuable tool to improve memory and learning. Future research should focus on the structure of procedures, and how the steps interact, rather than the quantity of steps required to follow the procedure to investigate the relationship between procedural complexity and the spacing effect.

Acknowledgments

Declarations

Funding

Aidan J. Horner is funded by the Economic and Social Research Council (ESRC) (ES/X00791X/1) and Leverhulme Trust (RPG-2023-003). Silke M. Göbel is funded by the Economic and Social Research Council [grant numbers ES/W002914/1, ES/W005654/1, ES/X00824X/1]. Ewan Murray was supported by a departmental studentship from the University of York.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare.

Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki

Declaration and its later amendments or comparable ethical standards. The study was approved by the University of York Psychology departments ethics board.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

Particiapants gave consent for their anonymised data to be published and stored in a OSF repository.

Availability of data and materials

All datasets used in this article are available in the following OSF repository (https://osf.io/cex8g/).

Code availability

The hypotheses and main analyses were pre-registered. The remaining analyses are clearly marked as not preregistered or are in the exploratory analyses sections. The code to replicate all the analyses in this study and this manuscript are available in the following OSF repository (https://osf.io/cex8g/).

For the purposes of open access, the authors have applied a creative comments attribution (CC-BY) license to any author accepted manuscripts version arising henceforth.

References

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

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

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

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. In *Form, function, and style in instructional design: Emerging research and opportunities* (pp. 1–26). https://doi.org/10.4018/978-1-5225-9833-6.ch001

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

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In *Psychology of learning and motivation* (Vol. 53, pp. 63–147). Elsevier. https://doi.org/https://doi.org/10.1016/S0079-7421(10)53003-2

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

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

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

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

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

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. S. (2016). Spaced retrieval practice increases college students’ short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, *28*(4), 853–873. https://doi.org/10.1007/s10648-015-9349-8

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

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

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

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

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

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

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020b). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, *32*(1), 277295. https://doi.org/10.1007/s10648-019-09489-x

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

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, *28*(6), 684–706. https://doi.org/10.1080/20445911.2016.1181637

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

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

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

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

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

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

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

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

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

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

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

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

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

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