Spaced Mathematics Practice Improves Test Scores and Reduces Overconfidence

William G. Emeny[1]

Marissa K. Hartwig[2]

Doug Rohrer[2]

1 Mathematics Department, Wyvern College, Hampshire, United Kingdom

2 Department of Psychology, University of South Florida, Tampa, USA

Correspondence: Doug Rohrer: drohrer@usf.edu

Abstract
The practice assignments in a mathematics textbook or course can be arranged so that most of
the problems relating to any particular concept are massed together in a single assignment, or
these related problems can be distributed across many assignments – a format known as
spaced practice. Here we report the results of two classroom experiments that assessed the
effects of mathematics spacing on both test scores and students' predictions of their test scores.
In each experiment, students in Year 7 (11-12 years of age) either massed their practice into a
single session or divided their practice across three sessions spaced one week apart, followed
one month later by a test. In both experiments, spaced practice produced higher test scores
than did massed practice, and test score predictions were relatively accurate after spaced
practice yet grossly overconfident after massed practice.

Keywords: mathematics; space; distribute; practice; confidence

Mathematics proficiency is improved by solving practice problems, but most mathematics assignments have features and characteristics that have proven inferior in randomized studies. Here we focus on one simple feature of mathematics practice: the extent to which practice problems relating to the same skill or concept are distributed throughout a course or textbook. For example, the practice problems in a textbook or course can be arranged so that most of the problems relating to hyperbolas are *massed* into a single assignment or *spaced* across many assignments. In the present research, we measured the effects of spaced practice on test scores in two classroom experiments with young mathematics students.

Our studies also examined the effect of spacing on students' *predictions* of their test scores. Students are notoriously overconfident, and overconfidence can lead students to make poor decisions about how and when they study. For example, they may stop practicing even though further practice would benefit them, or they might fail to seek clarification of concepts they do not understand. Overconfidence also can lead students (and their teachers) to believe that a chosen learning strategy is more effective than it actually is. For these reasons, it is important to know how confidence is affected by spacing and massing. In the experiments reported here, students predicted their test scores twice – once immediately after the last practice problem, and again immediately before the test. To our knowledge, no previous studies of mathematics learning have assessed the effect of spacing on students' judgments of their learning.

The Spacing Effect

Numerous studies have shown that distributing a fixed amount of practice over multiple sessions can boost scores on a delayed test – a finding known as the *spacing effect* (for recent reviews, see Carpenter, 2017; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Kang, 2016). Fewer studies have examined the effect of spacing on mathematics learning, though the available data suggest that spacing can boost mathematics learning, too. Mathematics spacing effects were first found in laboratory studies with college students (Gay, 1973; Rohrer & Taylor, 2006; Rohrer & Taylor, 2007), though one recent lab experiment found only null effects

(Ebersbach & Barzagar Nazari, 2020a). Spacing also improved test scores in each of several non-randomized classroom studies, including studies with students in high school geometry (Yazdani & Zebrowski, 2006), college statistics (Budé, Imbos, van de Wiel, & Berger, 2011), and year four mathematics (Chen, Castro-Alonso, Paas, & Sweller, 2018). Only in the last several years, however, has mathematics spacing been the focus of randomized studies in the classroom, and these too found test score benefits. These include a study with third- and seventh-grade students (Barzagar Nazari & Ebersbach, 2019), two studies fully embedded within a college pre-calculus course (Hopkins, Lyle, Hieb, & Ralston, 2016; Lyle, Bego, Hopkins, Hieb, & Ralston, 2020), and a study in a college statistics course (Ebersbach & Barzagar Nazari, 2020b).

Only a few studies of mathematics spacing have failed to find test benefits, and we suspect that these null effects reflect one or more boundary conditions. For instance, one of the laboratory studies cited above found that the spacing effect disappeared entirely when the test delay was shortened from four weeks to one week, suggesting that short test delays might reduce or eliminate the spacing effect (Rohrer & Taylor, 2006). Yet a recent laboratory experiment found null effects after both a one-week and five-week test delay, leading the authors to speculate that the nature of the mathematics task might moderate the size of the spacing effect (Ebersbach & Barzagar Nazari, 2020a). The present studies were not designed to test the plausibility of any particular boundary condition, but the results of our studies are incidentally inconsistent with certain possibilities, as we detail in the Discussion. That said, the present studies were designed to assess the effect of mathematics spacing on test scores in a classroom setting.

Students' Predictions of Test Performance

The second aim of the present work was to examine how mathematics spacing affects students' judgments of their future performance. Students who can accurately predict their future test performance are better able to make study decisions that are appropriate for their

current state of learning (Butler & Winne, 1995; Dunlosky & Rawson, 2012; Thiede, 1999), but previous studies have shown that students' predictions are often poorly calibrated (for a review, see Hacker, Bol, & Keener, 2008). Poor calibration has been observed in various subject areas – including mathematics (e.g., Barnett & Hixon, 1997; Hartwig & Dunlosky, 2017). A common form of poor calibration is overconfidence. In other words, when students are asked to predict their test performance before taking a test, their predictions often exceed their actual scores.

Techniques for improving students' prediction accuracy have been explored, with mixed success (Hacker, Bol, & Keener, 2008). While some techniques have involved direct metacognitive training, others have relied on students' own experiences with practice. For example, some strategies known to boost students' learning – such as practice testing – may be doubly beneficial because they can also improve prediction accuracy (Little & McDaniel, 2015). Whether the spacing of practice can also enhance prediction accuracy is unclear. Laboratory studies in which participants judged their learning following spaced or massed practice have yielded mixed results with regards to judgment accuracy (e.g., Kornell, 2009; Logan, Castel, Haber, & Viehman, 2012). Importantly, however, the possible benefit of spaced practice on prediction accuracy has not previously been examined in real classrooms or with mathematics materials.

Indeed, it is intuitive that a greater degree of spaced practice could promote more accurate predictions in classroom settings. When practice is spaced across multiple sessions, students will likely forget some of the to-be-learned material during the intervening time intervals, which in turn might draw their attention to the forgetting that often occurs across time (Koriat, Bjork, Sheffer, & Bar, 2004). Moreover, forgetting between consecutive sessions can help students recognize that their initial choice of learning strategy was suboptimal (cf. Bahrick & Hall, 2005) or that they had not learned the material as well as they believed. Massed practice, in contrast, may contribute to predictions that are overconfident because practice that occurs in a single, massed session may produce a high level of fluency or success within that session, without

signaling whether the learning will hold up across time. Put another way, students may fail to recognize that the gains experienced during massed practice will not persist as well as the gains experienced during spaced practice, resulting in an *illusion of mastery* (Bjork, 1999; Bjork, Dunlosky, & Kornell, 2013). To test these speculations about the effects of spaced and massed practice on prediction accuracy, students in the present studies predicted their future test score immediately after completing spaced or massed mathematics practice and again immediately before the test.

Overview of Present Research

In each of two studies, students completed 12 practice problems that were either massed in a single session or distributed evenly across three sessions spaced one week apart. Students were tested one month later. Students predicted their test scores twice: immediately after the last practice problem and immediately before the test. The studies were designed to answer two research questions: How does spacing affect test scores, and how does spacing affect students' predictions of their test scores?

Method

The methodology of the two experiments was nearly identical, and the slight differences are noted below. We chose to conduct Experiment 2 to replicate the results of Experiment 1 and to correct a potential issue concerning the mathematics proficiency of students in Experiment 1, as explained below. The anonymized data and all materials are available at Open Science Framework (OSF) https://osf.io/vcf6e/.

Students

Participating students attended a large school in the United Kingdom. Students completed the experiment during Year 7 (ages 11-12 years). We conducted Experiment 1 in spring 2018 and Experiment 2 in spring 2019. Experiment 1 was completed by 44 students, and Experiment 2 was completed by 55 students. These sample sizes were large enough to detect an effect size greater than $d = 0.43$ in Experiment 1 and $d = 0.38$ in Experiment 2 (assuming within-subjects

design, two-tailed test, alpha = .05, and power = 0.8). An additional 17 students in the participating classes (11 in Experiment 1, and 6 in Experiment 2) missed at least one of the class meetings in which the experiment took place, and their incomplete data were excluded from all analyses.

Each experiment included two participating classes, which we arbitrarily label Class A and Class B. In Experiment 1, Class A included 24 participating students (12 girls, 12 boys), and Class B had 20 (10 girls, 10 boys). In Experiment 2, Class A included 28 participating students (11 girls, 17 boys), and Class B had 27 (13 girls, 14 boys). We measured students' mathematics proficiency with the KS2 SAT mathematics score, which we were able to obtain for all but three participating students (one in Experiment 1, and two in Experiment 2). In Experiment 1, the KS2 SAT scores were significantly greater in Class A ($M = 109.6$, $SD = 3.4$, $n = 23$) than in Class B ($M = 104.4$, $SD = 4.4$, $n = 20$), $p < .001$, $d = 1.33$. In Experiment 2, there was no reliable difference between the scores of Class A ($M = 110.6$, $SD = 4.8$, $n = 27$) and Class B ($M = 108.8$, $SD = 5.0$, $n = 26$).

Materials and Design

Each mathematics problem included a Venn diagram task or a permutation task (Figure 1). Class A spaced their practice of the Venn problems and massed their practice of the permutation problems, and Class B did the reverse. We created the Venn problems, and the permutation task was taken from Rohrer and Taylor (2006). We did not administer a pretest, though neither task is part of the school curriculum or national curriculum for Year 7 or sooner.

Procedure

The timeline is shown in Figure 2. For both the Venn and permutation tasks, students saw a tutorial followed by 12 practice problems that were either massed into a single session or divided evenly across three practice sessions spaced one week apart. The test delay equaled four weeks (27-29 days). The sessions were scheduled so that students never completed a task for both the spaced and massed condition on the same day. Instead, the timelines for the two

conditions (spaced and massed) were staggered so that corresponding sessions for the two conditions were offset by one or two days. Finally, we counterbalanced condition and test order so that Class A was tested on the spaced material before they were tested on the massed material, and Class B was tested on the massed material before they were tested on the spaced material.

Students completed the practice sessions and tests during their mathematics class under the supervision of their teacher. Neither of the two teachers was a member of the research team, and no researcher attended a class meeting in which the experiment took place. Students solved problems by paper and pencil, without a calculator. No problem appeared more than once during the experiment. For each task (Venn and permutation), students saw a tutorial immediately before they began the first practice problem, and both the tutorial and practice problems were presented via an audiovisual slideshow created by the first author. Each practice problem appeared one at a time. Each problem remained on the screen while students attempted to solve it (45 s). Immediately afterwards, the solution appeared on the screen and remained on the screen while students heard a prerecorded oral explanation (60 s, on average), and then the solution remained on the screen while students corrected their errors (20 s). Having students correct their solution immediately after each practice problem prohibited us from scoring their practice problems, and thus we have no measure of practice performance. We should also note that this immediate feedback ensured that the study manipulated not only practice but instruction more broadly. Throughout each practice session, students could see their written work for practice problems that had appeared previously in the session. For each test (Venn and permutation), students were given a sheet of paper listing four novel problems and asked to solve the problems in any order they wish (8 min). Students received no feedback during the test.

Students predicted their test scores twice: immediately after the last practice problem, and immediately before the test. They also "postdicted" their test score immediately after the test.

For each of these three judgments, students were asked to estimate their score on a test with four problems, as detailed in the Appendix.

<div align="center">**Results**</div>

Test Scores

Both experiments showed a spacing effect (Figure 3). In Experiment 1, the spacing effect was moderately large, $t(43) = 4.0$, $p < .001$, $d = 0.61$, 95% CI [0.28, 0.92]. In Experiment 2, the spacing effect was smaller but still sizeable, $t(54) = 2.9$, $p < .01$, $d = 0.39$, 95% CI [0.11, 0.66]. Each Cohen's d value equals the mean difference score divided by the standard deviation of the difference scores (also known as d_z). This choice of Cohen's d produced a smaller value than other alternatives because the two sets of scores were not strongly correlated, and thus the reported effect sizes are conservative (Lakens, 2013).

The spacing effect was not associated with student proficiency. Specifically, students' mathematics scores on the KS2 SAT (see Method) were not significantly correlated with the students' spacing effect (test score in spaced condition minus test score in massed condition) in both Experiment 1 ($r = .21$, $p = .19$) and Experiment 2 ($r = .07$, $p = .63$). Thus, the data provide no support for the claim that the benefits of spacing depend on student proficiency.

Students' Predictions

In both experiments, spaced practice led to accurate predictions of test scores, whereas massed practice engendered overconfidence (Figure 4). For Prediction 1, which students made immediately after the last practice problem, spacing led to predictions that were slightly but not significantly greater than actual test scores (Experiment 1: $t(43) = 1.0$, $p = .33$, $d = 0.15$; Experiment 2: $t(54) = 1.7$, $p = .10$, $d = 0.23$), whereas massed practice led to substantial overconfidence (Experiment 1: $t(43) = 6.1$, $p < .001$, $d = 0.92$; Experiment 2: $t(54) = 7.9$, $p < .001$, $d = 1.07$). For Prediction 2, which students provided immediately before the test, spacing again led to accurate predictions (Experiment 1: $t(43) = -0.1$, $p = .93$, $d = 0.01$; Experiment 2: $t(54) = 0.1$, $p = .93$, $d = 0.01$), whereas massed practice produced moderate overconfidence

(Experiment 1: $t(43) = 2.7$, $p < .01$, $d = 0.41$; Experiment 2: $t(54) = 2.3$, $p = .03$, $d = 0.31$).

Finally, the "postdictions" made immediately after the test were accurate regardless of whether practice had been spaced or massed.

Because the accurate predictions of test scores necessarily yield a null difference between the predicted test score and the actual test score, we also found the Bayes Factor for every null difference reported above to determine whether the evidence favors the null hypothesis (accuracy) or the alternative hypothesis (inaccuracy). We used an online algorithm created by Rouder, Speckman, Sun, Morey, and Iverson (2009). The algorithm assumes a Jeffreys prior, and we set scale r to the default value of 0.7071. In every case, the Bayes Factor exceeded one, which means that the evidence did in fact favor the null hypothesis (accuracy).

Discussion

Spaced mathematics practice sharply improved test scores in both experiments (Figure 3). In addition, students' predictions of their test scores were quite accurate if they had spaced their practice yet grossly overconfident if they had massed their practice (Figure 4). Both findings have caveats and practical implications.

Boundary Conditions of the Mathematics Spacing Effect

Although mathematics spacing effects have been observed repeatedly, a few studies have found null effects (see Introduction). An examination of the data suggests that these negative findings might be due to one or more boundary conditions. Here we consider four possibilities.

1. *Nature of the Task*. Might the mathematics spacing effect be moderated by the nature of the mathematics task? This possibility was raised by Ebersbach and Barzagar Nazari (2020a) to account for the null effects observed in every condition of their laboratory study. Yet we are not aware of any evidence for this claim. The study by Ebersbach and Barzagar Nazari included only one kind of task, and it was the same permutation task that produced large spacing effects in each of two previous studies (Rohrer & Taylor, 2006; 2007). Furthermore, no studies have

shown an association between the nature of a mathematics task and the size of the spacing effect.

2. *Short Test Delay*. Some evidence suggests that the mathematics spacing effect can vanish when the test delay is brief. In one laboratory study, students either massed their practice or spaced their practice across two sessions spaced one week apart before taking an exam one or four weeks later, and only the four-week delay produced a spacing effect (Rohrer & Taylor, 2006). Similar findings have been obtained in spacing studies with non-mathematics materials (e.g., Rawson & Kintsch, 2005; Serrano & Muñoz, 2007). Findings such as these suggest that mathematics test scores might not benefit from spacing if test delays are brief.

3. *Short Spacing Gaps.* Some indirect evidence suggests that the mathematics spacing effect can shrink or disappear if the spacing gap is much shorter than the test delay, as is the case with non-mathematics materials (e.g., see meta-analysis by Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). This possibility also is consistent with the results of the mathematics spacing study that found null effects when the test delay was rather long (Ebersbach & Barzagar Nazari, 2020a). Students in that study either massed their practice or spaced their practice across two sessions separated by 1 or 11 days before taking a test five weeks later, and the larger spacing gaps produced larger test scores: about 25% after massed practice, about 30% after a 1-day spacing gap, and nearly 40% after an 11-day spacing gap. Thus, although the spacing effects were not statistically significant, both spacing gaps produced positive spacing effects, and the effect size was smaller for the 1-day gap than for the 11-day gap. If too-short spacing gaps do reduce the size of the spacing effect, mathematics students should space their practice across sessions separated by at least a week or more if long-term learning is the goal.

4. *Sparse Feedback.* The test score benefits of spaced mathematics practice might fade if students are not promptly shown the correct solution to any problem they cannot solve. In the present studies, students were shown the correct solution to each problem immediately after each attempt, and they also were required to correct any errors in their solution. Without this

kind of instructional feedback, spaced practice can be a disadvantage because students who forget how to solve a particular kind of problem during the spacing gap will then be unable to either solve the problem or learn from feedback, thereby denying students an opportunity to learn during this subsequent session. This potential caveat is only conjecture, however, because the nature or degree of feedback has not been manipulated in a mathematics spacing study. However, feedback was provided after only one half of the practice problems in the aforementioned study finding only null effects (Ebersbach & Barzagar Nazari, 2020a).

Student Overconfidence

Apart from the spacing effect, the present studies showed that spacing can improve students' accuracy in judging their learning. Following spaced practice, students predicted their future test scores very accurately, whereas massed practice yielded gross overconfidence. The overconfidence after massed practice might be due to the fluency or success with which students can solve a set of similar problems by merely repeating the same procedure over and over, giving the impression that students have mastered the content. In turn, overconfidence may lead students and their teachers to believe that further practice is unnecessary when, in fact, the gains will not be retained across time.

Though massed practice produced overconfidence, we note that predictions following massed practice were only slightly greater than predictions following spaced practice. Thus, massed practice did not elevate predictions to an unrealistically high level but instead failed to help students recognize their low level of mastery. One might speculate that awareness could be improved by assigning practice problems after a delay – essentially a practice test – which can boost both metacognitive awareness and test scores. Spaced mathematics practice provides exactly this kind of practice testing. With spaced practice, both students and teachers get a clearer picture of students' comprehension and retention across time and are better poised to steer future practice effectively.

Practical Implications

    The learning benefits of spaced mathematics practice have obvious utility in the classroom. Spacing improved test scores in the present studies by more than half, which is far more beneficial than most mathematics learning interventions. These results, in conjunction with the results of other researchers, provide strong support for spaced mathematics practice. More broadly, we believe spacing improves the long-term learning of any kind of material, and we join the many other researchers who have previously advocated for a greater degree of spacing in the classroom (e.g., Carpenter, 2017; Kang, 2016; Dunlosky et al., 2013, Roediger & Pyc, 2012; Willingham, 2014).

    Still, spaced mathematics practice is not without its challenges. Apart from the possible boundary conditions described above, little is known about exactly how practice problems should be spaced. For instance, a dozen practice problems can be distributed evenly across three assignments or distributed more thinly across six assignments, and the spacing gaps between assignments can be fixed in length (always one week) or expanding (1 week, 3 weeks, and then 9 weeks). The variations are uncountable. Finally, spaced mathematics assignments are not readily available to many teachers because students' mathematics textbooks provide only a small degree of spacing (Rohrer, Dedrick, & Hartwig, 2020). These teachers would need to create their own spaced assignments or use spaced assignments drawn from the internet or other resources.

    Yet these obstacles should not overshadow the merits of spaced mathematics practice. Spacing is one of the largest and most robust learning strategies known to learning researchers, and spacing can be implemented in nearly any mathematics course. Though there are some subtleties regarding its implementation, the key point is that teachers should shift their mindset so that the practice of a skill or concept is seen not as material that should be squeezed into one or two consecutive class meetings but rather as material that can be distributed across

many lessons. By adopting this approach, mathematics teachers can help their students better

learn the material and also better gauge how well they have learned the material.

References

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

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*, 170–174.

Barzagar Nazari, K., & Ebersbach, M. (2019). Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, *33*, 288–298.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher and A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

Budé, L., Imbos, T., van de Wiel, M. W., & Berger, M. P. (2011). The effect of distributed practice on students' conceptual understanding of statistics. *Higher Education, 62*, 69–79.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281.

Carpenter, S. K. (2017). Spacing effects in learning and memory. In J. T. Wixted (Ed.), *Cognitive Psychology of Memory,* Vol. 2 Learning and Memory: A Comprehensive Reference, 2nd edition, J. H. Byrne (Ed.), pp. 465–485. Oxford: Academic Press.

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*, 354–380.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect. *Educational Psychology Review*, *30*, 483-501.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280.

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

Ebersbach, M., & Barzagar Nazari, K. (2020a). No robust effect of distributed practice on the short- and long-term retention of mathematical procedures. *Frontiers in Psychology*, *11*, 811.

Ebersbach, M., & Barzagar Nazari, K. (2020b). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *Journal of Applied Research in Memory and Cognition*, *9*, 532-541.

Gay, L. R. (1973). Temporal position of reviews and its effect on the retention of mathematical rules. *Journal of Educational Psychology, 64,* 171–182.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York: Taylor & Francis.

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80–90.

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. (2016). Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge. *Educational Psychology Review*, *28*, 853–873.

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133, 643–656.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297–1317.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Little, J. L., & McDaniel, M. A. (2015). Metamemory and control following retrieval practice for text. *Memory & Cognition, 43*, 85–98.

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

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

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology, 97,* 70–80.

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242–248.

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in

mathematics textbooks. *Educational Psychology Review*, *32*, 873-883.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science, 35,* 481–498.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System, 35*, 305–321.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin and Review, 6*, 662–667.

Willingham, D. T. (2014). Strategies that make learning last. *Educational Leadership*, *72(2),* 10–15.

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*, 37–43.

Appendix

Students were asked to predict their test scores at three points in time. The prompts for each prediction are listed below, and complete materials can be found at OSF https://osf.io/vcf6e/.

Prediction 1 (immediately after the last practice problem). In 4 weeks' time you are going to take a 4 question test on this topic. How many questions out of 4 do you think you will get correct on that test?

Prediction 2 (immediately before the test). You are about to take a 4 question test on the topic of [permutations/Venn diagrams]. This will be similar to the practice questions you answered several weeks ago which looked like this: [sample problem like shown in Figure 1]. How many questions out of 4 do you think you will get correct on this test today?

Postdiction (immediately after the test). How many questions out of 4 do you think you got correct on the test you have just taken?
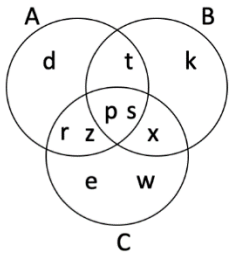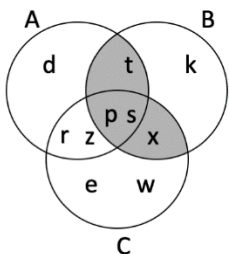
Figure 1

Permutation	Venn
Problem	**Problem**
How many different ways can the following letters be arranged?	The Venn Diagram shows the elements of sets A, B and C. List all the elements of
AABCC	$(A \cup C) \cap B$
Solution	**Solution**
$$\frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = \frac{60}{2} = 30$$	{t, p, s, x}

Figure 1. The tasks. Every student saw the same practice problems, and no student saw the same problem twice. Immediately after each practice problem, students were shown the solution and asked to correct any errors.

Figure 2

Spaced condition   Massed condition

Tutorial
4 problems

1 wk

4 problems

1 wk   Tutorial

4 problems   12 problems
Prediction 1   Prediction 1

4 wks   4 wks

Prediction 2   Prediction 2
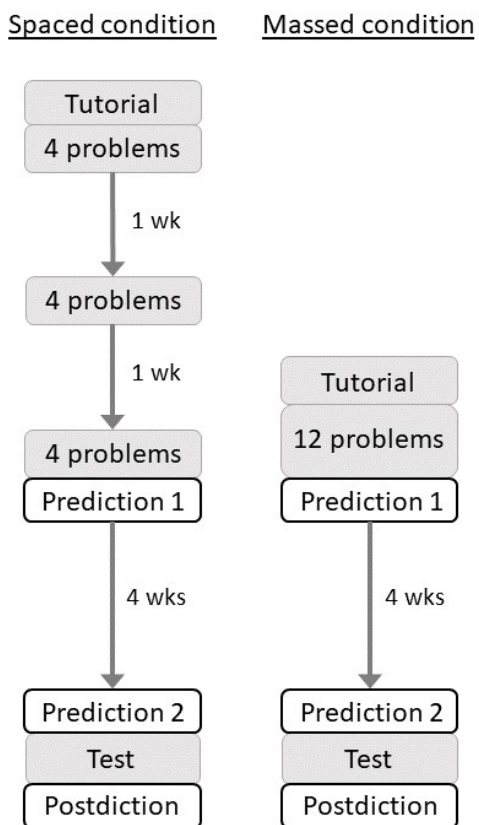Test   Test
Postdiction   Postdiction

Figure 2. The procedure. Students completed 12 practice problems. Students predicted their test scores immediately after the last practice problem (Prediction 1) and immediately before the test (Prediction 2). Students also estimated their test scores immediately after the test (Postdiction).
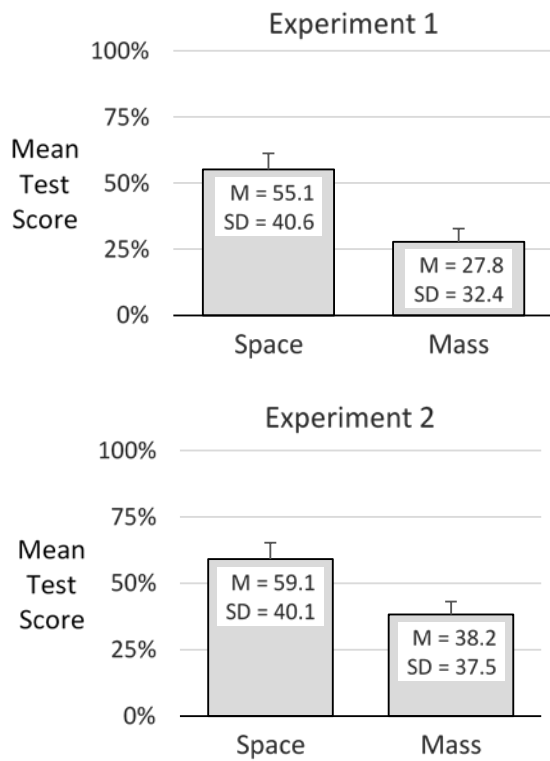
Figure 3

Figure 3. Test scores. Both studies showed a spacing effect. Error bars equal one standard error.

Figure 4

Experiment 1

Mean Test Score (max = 4)

Space

Mass

Test Score

Test Score

Prediction 1 Prediction 2 Postdiction

Prediction 1 Prediction 2 Postdiction

Experiment 2

Mean Test Score (max = 4)

Space

Mass

Test Score

Test Score

Prediction 1 Prediction 2 Postdiction

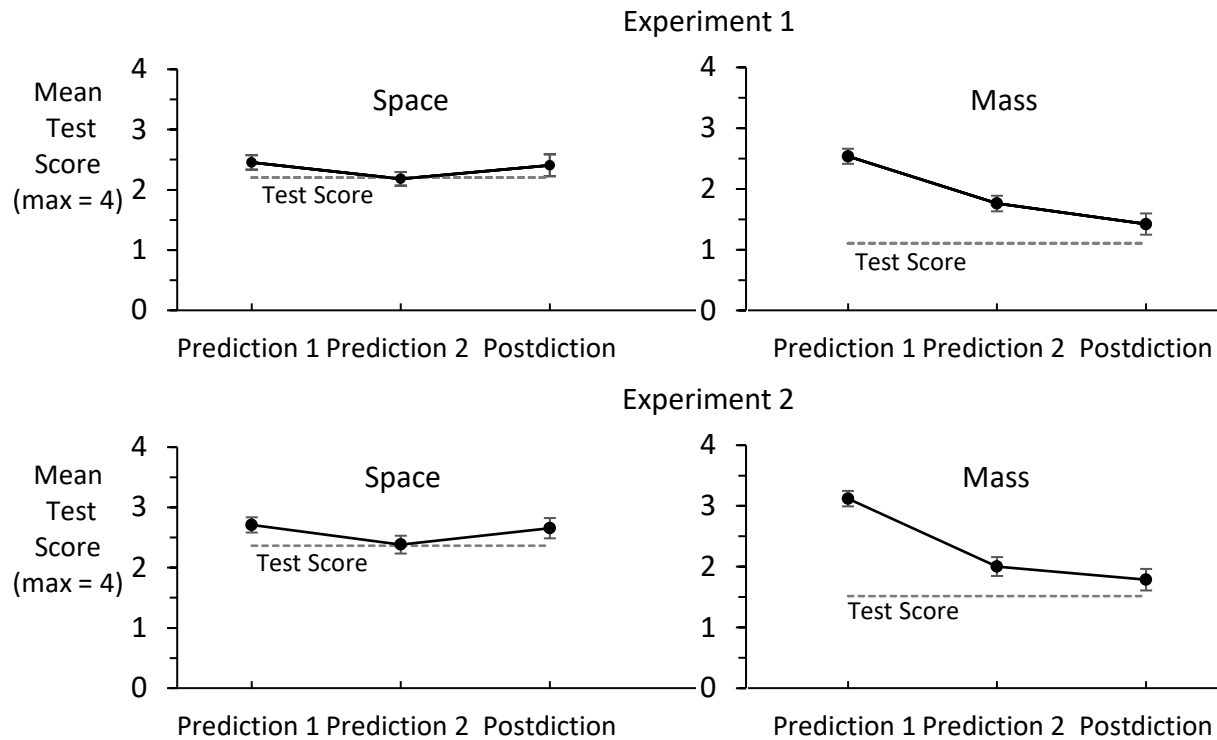Prediction 1 Prediction 2 Postdiction

Figure 4. Predicted test scores. Massed practice led to overconfidence at Prediction 1 (immediately after the last practice problem) and Prediction 2 (immediately before the test). Students were asked to estimate the number of problems (out of four) they would solve correctly (see Appendix). Dashed lines indicate the actual test scores, which are also shown in Figure 3. Error bars equal plus or minus one standard error.