International Journal of STEM Education

RESEARCH

Open Access

Single-paper meta-analyses of the effects of spaced retrieval practice in nine introductory STEM courses: is the glass half full or half empty?

Campbell R. Bego[1*], Keith B. Lyle[2], Patricia A. S. Ralston[1], Jason C. Immekus[3], Raymond J. Chastain[4], Lora D. Haynes[2], Lenore K. Hoyt[5], Rachel M. Pigg[6], Shira D. Rabin[6], Matthew W. Scobee[7] and Thomas L. Starr[8]

Abstract

Background Undergraduate STEM instructors want to help students learn and retain knowledge for their future courses and careers. One promising evidence-based technique that is thought to increase long-term memory is spaced retrieval practice, or repeated testing over time. The beneficial effect of spacing has repeatedly been demonstrated in the laboratory as well as in undergraduate mathematics courses, but its generalizability across diverse STEM courses is unknown. We investigated the effect of spaced retrieval practice in nine introductory STEM courses. Retrieval practice opportunities were embedded in bi-weekly quizzes, either massed on a single quiz or spaced over multiple quizzes. Student performance on practice opportunities and a criterial test at the end of each course were examined as a function of massed or spaced practice. We also conducted a single-paper meta-analysis on criterial test scores to assess the generalizability of the effectiveness of spaced retrieval practice across introductory STEM courses.

Results Significant positive effects of spacing on the criterial test were found in only two courses (Calculus I for Engineers and Chemistry for Health Professionals), although small positive effect sizes were observed in two other courses (General Chemistry and Diversity of Life). Meta-analyses revealed a significant spacing effect when all courses were included, but not when calculus was excluded. The generalizability of the spacing effect across STEM courses therefore remains unclear.

Conclusions Although we could not clearly determine the generalizability of the benefits of spacing in STEM courses, our findings indicate that spaced retrieval practice could be a low-cost method of improving student performance in at least some STEM courses. More work is needed to determine when, how, and for whom spaced retrieval practice is most beneficial. The effect of spacing in classroom settings may depend on some design features such as the nature of retrieval practice activities (multiple-choice versus short answer) and/or feedback settings, as well as student actions (e.g., whether they look at feedback or study outside of practice opportunities). The evidence is promising, and further pragmatic research is encouraged.

Keywords Spaced retrieval practice, STEM education, Applied research, Single-paper meta-analysis

*Correspondence:
Campbell R. Bego
campbell.bego@louisville.edu
Full list of author information is available at the end of the article

Springer Open

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Background

The term "STEM education" has enjoyed a roughly 20-fold increase in usage from 2000 to 2019 (Google nGram), exemplifying the growing societal concern with educating future professionals in the disciplines of science, technology, engineering, and mathematics (STEM). Ideally, learning scientists would meet this concern by identifying pedagogical innovations that can be broadly deployed to improve educational outcomes in courses in all the STEM disciplines. From a curriculum administration perspective, it is much easier to adopt a single, robust innovation rather than to tailor change, course-by course. On the one hand, this seems possible, given that the umbrella term STEM implies some degree of similarity between the constituent fields and perhaps some shared reactivity to educational interventions. On the other hand, STEM courses are far from monolithic, and their myriad differences might prevent techniques that work in one course from generalizing to others. Ultimately, only empirical research can determine whether any particular technique is broadly effective across courses in multiple STEM domains. In the present research, we studied the efficacy of a technique known as *spaced retrieval practice*, which has previously been shown to increase knowledge retention in a variety of settings. Here, we tested the technique's effectiveness simultaneously in several different undergraduate STEM courses. In this study, we sought to examine whether and how spaced retrieval practice would enhance memory across STEM courses, irrespective of discipline.

Retrieval practice + spacing: a potential solution to an age-old problem

Some common undergraduate studying practices, such as rereading notes and textbooks, promote only fleeting memory of course content. Information is retained long enough to support adequate performance on near-term tests but is forgotten soon after (Bacon & Stewart, 2006; Conway et al., 1991; Kamuche & Ledman, 2005; Rawson et al., 2013). Forgetting course content is antithetical to the central premise of education and is especially problematic when success in higher-level courses depends on retention from lower-level courses. Undergraduates pursuing degrees in STEM fields often need the knowledge they acquired in early courses to understand content in advanced courses and to perform in the jobs they hope to obtain.

Basic principles of learning and memory suggest ways to modify educational practices that may make them more likely to foster long-term knowledge retention. Cognitive psychologists have weighed in on the value of various modifications (Dunlosky et al., 2013; Pashler et al., 2007; Roediger & Pyc, 2012), basing their recommendations on a large body of laboratory experiments and a smaller number of classroom experiments. One technique that has been especially recommended is spaced retrieval practice.

The term *retrieval practice* refers to the repeated retrieval of a piece of information, typically in preparation for a future test. Effects of retrieval practice on memory have been extensively studied (Rowland, 2014). The robust finding is that retrieving information from memory bolsters the long-term retention of that information more than restudying the information (i.e., the testing effect; e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006; for a meta-analysis, see Adesope et al., 2017; for a recent review, see McDermott, 2021). In other words, if you want to remember something over the long term, you are better served by repeatedly retrieving the information (e.g., by answering questions) than by repeatedly restudying the information (e.g., by rereading). College students often fail to harness the power of retrieval practice, choosing instead to restudy (Karpicke et al., 2009). Classroom studies have shown that, when instructors implement assignments or activities that require students to practice retrieval more than students would on their own, students retain more course content and earn higher grades (e.g., Leeming, 2002; Lyle & Crawford, 2011; McDaniel et al., 2007; for a review, see Agarwal et al., 2021; Yang et al., 2021).

Retrieval practice is said to be *spaced* when instances of retrieving the same piece of information are spread out over time. In contrast, it is said to be *massed* when multiple retrievals occur consecutively or in a brief temporal window. These terms are often used in a relative sense, such that any two conditions in which retrievals are more spaced out in one than the other are called the spaced and massed conditions, respectively. When spaced and massed retrieval conditions are compared in laboratory studies with verbal materials, the spaced condition typically leads to superior information retention (Cepeda et al., 2006; Cull, 2000; Karpicke & Roediger, 2007; Landauer & Eldridge, 1967), representing what is called a *spacing effect* (Dempster, 1989).[1] A small number of laboratory studies have examined mathematics knowledge and found that it also benefits from spaced, versus massed, retrieval practice (Rohrer & Taylor, 2006, 2007).

Common educational practices seemingly promote massed retrieval practice more than spaced, including in STEM domains. For example, an analysis of mathematics textbooks showed that practice problems are

[1] Historically, the term lag effect was used when a comparison was made between two conditions with more and less spacing, whereas the term spacing effect was reserved for comparing a condition with spacing to a condition without spacing (Kahana & Howard, 2005).

overwhelmingly massed (Rohrer et al., 2020). More generally, massed retrieval practice tends to go hand in hand with the ubiquitous practice of separating educational content into units, since opportunities to retrieve content rarely extend beyond the unit in which it is presented. As for students themselves, they seem not to be cognizant of the value of spacing (Hartwig et al., 2022; Kornell, 2009; Logan et al., 2012), making it unlikely that they will spontaneously incorporate spaced retrieval into their own self-initiated study practices. All of this suggests that incorporating more spaced retrieval practice into classrooms could greatly increase students' long-term retention of vital course content. In addition, spaced retrieval practice may indirectly benefit learning by improving students' metacognition or self-regulated learning processes (e.g., Ariel & Karpicke, 2018).

When considering the classroom application of spaced retrieval practice, it is important to consider the effect of spacing not only on long-term retention but also on the retrieval practice exercises themselves. In some cases, spacing has been shown to reduce performance on retrieval practice exercises (e.g., Lyle et al., 2020). Indeed, spacing is often characterized as inducing *desirable difficulty* (Bjork, 1994), meaning that spacing reduces performance during the early stages of learning but, in the process, promotes cognitive mechanisms that benefit long-term retention (Bjork & Bjork, 2011; Bjork, 1999). In our reading of the literature, however, we have found it unclear whether spacing induces difficulty, either with or without long-term benefits, when learning educationally relevant material. Several studies in the STEM domain of mathematics and applied mathematics have provided relevant data. Of these, some have found a negative effect of spacing on the retrieval practice exercises (Ebersbach & Barzagar Nazari, 2020b; Lyle et al., 2020), but others have not (Barzagar Nazari & Ebersbach, 2019; Ebersbach & Barzagar Nazari, 2020a; Rohrer & Taylor, 2007).

Classroom research on spaced retrieval practice

Although cognitive psychologists have recommended using spaced retrieval practice to enhance memory in educational settings (Carpenter et al., 2012; Dunlosky et al., 2013; Kang, 2016; Roediger & Pyc, 2012; Weinstein et al., 2018), it is important to ask how much we know about its impact in actual courses. Classroom studies on spaced retrieval practice are rare compared to laboratory studies. For example, in a meta-analysis of research that was published in September 2017, Latimier et al. (2021) identified 29 articles that examined the effect of spaced versus massed retrieval practice on retention. Of those, only 3 were labeled as classroom studies and one of those appears to have been mislabeled (Storm et al., 2010). The remaining two showed significant positive effects of increasing the spacing of retrieval practice in undergraduate STEM courses (structural kinesiology in Dobson et al., 2017, and precalculus in Hopkins et al., 2016). A third study, seemingly mislabeled as a laboratory study in Latimier et al. (2021), showed a significant positive effect of spaced retrieval practice in a high school physics class (Grote, 1995). More recently, the positive effect of spaced retrieval practice was replicated in another undergraduate precalculus course (Lyle et al., 2020) and obtained in undergraduate calculus (Lyle et al., 2022) and statistics courses (Ebersbach & Barzagar Nazari, 2020a; see also Budé et al., 2011, for related research less tightly focused on retrieval practice), as well as in math courses for German third and seventh graders (Barzagar Nazari & Ebersbach, 2019). Less encouraging, spacing out retrieval practice had no clear-cut benefit for the retention of course content in introductory psychology courses (Burns & Gurung, 2023; Gurung & Burns, 2019).

Summarizing the available classroom research, it is, on balance, promising, but skewed toward mathematics and the applied mathematics field of statistics. Even if we accept that spaced retrieval practice works well in those domains, it is probably premature to assert that it works well in all classes. The many enthusiastic calls to embrace spacing in educational contexts (Carpenter et al., 2012; Dunlosky et al., 2013; Kang, 2016; Pashler et al., 2007; Roediger & Pyc, 2012; Weinstein et al., 2018) may be running ahead of the available evidence. Phenomena that seem clear and robust in the laboratory are not always readily discernible in the classroom (Fyfe et al., 2021). In the classroom, students can decide when and how to interact with course material, and this makes it difficult to detect effects that depend critically on the timing of events. Recommendations to implement spacing would be on firmer footing if based on data from classrooms across a variety of disciplines. One way to begin accumulating such evidence is to look for a spacing effect in courses in the STEM disciplines, because they are often grouped together, and because work has already been done in mathematics courses. If there were evidence that spaced retrieval practice is *not* generally effective across STEM courses, it might require revising recommendations, which, at present, do not take domain-specificity or other course variations into account.

It is necessary to consider the effectiveness of any educational intervention across contexts and populations to fully assess its real-world utility and to understand parameters that modulate its amplitude (Dunlosky et al., 2013). Fyfe et al. (2021) proposed the *ManyClasses* paradigm as a framework to assess the generalizability of an educational practice. In this paradigm, an experiment with a single research question and well-designed methodology is simultaneously implemented in many different

Table 1 Course names and disciplines

Course name	S. T. E. or M	Requisite for:
Chemical engineering thermodynamics	Engineering	Chemical Engineering
Chemistry for health professionals	Science	Nursing
Diversity of life	Science	Biology/Medicine
Fundamentals of physics I	Science	Physics/Medicine
General chemistry	Science	Chemistry/Engineering/Medicine
Research methods for psychology	Science	Psychology
Statistics for psychology	Math, Science	Psychology, Statistics-based STEM degrees
Unity of life	Science	Biology/Medicine
(Calculus I for Engineering Students)	(Mathematics, Engineering)	(Engineering)

Calculus I for Engineering Students was a course in this study, but spacing was implemented through a different online platform in this course, and thus the methods and results were published separately (Lyle et al., 2022)

courses. Some flexibility in the implementation is allowed to accommodate different course structures. Classrooms in such a study can vary along numerous dimensions including type of institution, instructor experience, course size, and student demographics. By studying versions of the intervention in many classes, researchers can test its effectiveness in authentic environments without limiting the scope of the results to a single context. This paradigm provides replication, variation, and ecological validity, all of which help contribute to the idea of generalizability. A main effect in a *ManyClasses* study indicates that an educational intervention is likely to be effective in many different classrooms.

Current study

The current study investigated spaced retrieval practice in ten different introductory STEM classes, in keeping with the goal of *ManyClasses*. This study therefore evaluates the benefits and potential costs of spacing in a variety of authentic educational contexts. Our specific research questions were: What is the effect of spaced retrieval practice on (RQ1) practice quizzes and (RQ2) long-term knowledge retention in a variety of STEM courses?

The full study originally comprised ten introductory STEM courses in six STEM disciplines (biology, chemistry, engineering, mathematics, physics, and psychology) at a mid-sized public research university. Unfortunately, the introductory algebra course had lower enrollment than anticipated and could not be included in our analyses. In addition, an engineering calculus course—the only other mathematics course in this study—was delivered through a different learning management system. Since all other courses used the same learning management system, the methods and results from calculus were published separately (Lyle et al., 2022). However, because calculus was part of our original study design, we incorporate the results in some of our analyses, as well as our discussion of the results.

The current manuscript therefore presents new results from eight introductory STEM courses, an analysis of the effect across these eight courses, and an analysis of the effect across nine courses with the addition of calculus. The course names, disciplines, and related degrees are shown in Table 1 below. Courses were selected that were required for various STEM degrees across a diversity of domains, with large class sizes and an instructor willing to participate. Although the current work consists of a relatively small number of courses compared to the number of classrooms in the seminal *ManyClasses* paper ($N=38$; Fyfe et al., 2021), we believe we have estimated generalizability beyond a single study, capturing the spirit of the *ManyClasses* initiative.

In each course, we implemented a within-subjects manipulation of the spacing of retrieval practice. All students retrieved knowledge about some topics in a spaced condition and retrieved knowledge about other topics in a massed condition. In addition to controlling for individual performance differences, the within-subjects study design was selected because it was equitable. All students experienced increased spacing for some of the course content and thus all students had the opportunity to benefit from this intervention.

The vehicle for this manipulation was a set of five practice quizzes administered over the course of a semester. Retention of learning objectives following spaced practice was compared to retention of learning objectives following massed practice on a criterial test at the end of the semester.

Table 2 Assessment item type by course

Course name	Multiple-choice		Calculate	Fill-in-the-blank
	Standard	Modified		
Chemical engineering thermodynamics	76		20	
Chemistry for health professionals	36	46		14
Diversity of life	76	20		
General chemistry	16	61	4	15
Fundamentals of physics I	96			
Research methods for psychology	96			
Statistics for psychology	29	44	4	19
Unity of life	96			

Based on previous findings in precalculus and calculus classrooms (Hopkins et al., 2016; Lyle et al., 2020, 2022), we hypothesized that spaced retrieval practice would increase knowledge retention in at least some of the classes studied here. We did not necessarily expect spacing to produce statistically significant results in all classes, however, because spaced retrieval had not been tested previously in a variety of STEM classrooms. Our goal was to identify both the classes in which spacing enhanced performance and the classes in which it did not.

Method

This research was approved as an exempt study with waiver of informed consent by the Institutional Review Board at the university where it was conducted.

Participants

Participants were 910 undergraduate students enrolled in one[2] of the eight introductory STEM courses involved in this research. Participants who experienced technical problems resulting in additional retrievals ($N=38$) and those who failed to complete all experimental phases of the research ($N=294$) were excluded from analyses. The total number of retained participants was 578. Gender and racial demographics can be found in the supplemental information (Additional File 1). Sample sizes were as follows: Chemical Engineering Thermodynamics, $N=42$; Chemistry for Health Professionals, $N=112$; Diversity of Life, $N=51$; Fundamentals of Physics I, $N=106$; General Chemistry, $N=61$; Research Methods for Psychology, $N=30$; Statistics for Psychology, $N=74$; and Unity of Life, $N=102$.

[2] A small number of participants (< 6%) were enrolled in more than one course in this study. We conducted analyses both including and excluding dual-enrollment participants. Results were similar, and thus only analyses all participants are presented in this manuscript. All data are available online for review.

Materials

This section provides information about materials that pertains to all courses. The full set of materials for each course including calculus can be found in the supplemental online resources (Additional Files 2, 3, 4, 5, 6, 7, 8, 9, 10, 11).

Materials consisted of five quizzes and a criterial test in each of the eight STEM courses. Quiz and test questions were designed to assess student knowledge of 24 specific learning objectives (henceforth called target learning objectives), drawn from the larger pool of learning objectives in each course. Only objectives introduced in the first seven weeks of the semester were candidates for selection because there had to be sufficient time in the semester following an objective's introduction to space retrieval practice across several weeks. In every course, instructors selected eight objectives from weeks 1–3, eight from weeks 4–5, and eight from weeks 6–7. All objectives were deemed to be fundamental topics in the courses. Each learning objective was assessed with four items, for a total of 96 assessment items per course.

Although asking the same question multiple times in one quiz is common in laboratory studies, this would be a strange practice in a classroom setting. Therefore, in this study, four items were created that differed in verbiage but required students to retrieve the same information (see Bego et al., 2020). The research team and instructors worked carefully together to construct and review all items prior to implementation. All assessment items are available for review in the supplemental resources.

Item types consisted of (a) standard multiple-choice questions (i.e., two to four possible responses with one correct answer); (b) modified multiple-choice questions (e.g., multiple correct responses, sorting or classifying multiple items, or answer lists with more than five possible response options like numbers 0–9); (c) calculation questions requiring a numerical response; and (d) fill-in-the-blank responses. The prevalence of different question

Standard Multiple-Choice Question from *Unity of Life*:

Polymers are broken down with _____ reactions.

*a. hydrolysis
b. dehydration
c. both, depending on the polymer

Modified Multiple-Choice Question from *Statistics for Psychology*:

Which of these is an inferential statistical technique?

a. Find the mode
*b. Analyze data using a one-way ANOVA
*c. Conduct a factorial ANOVA
*d. Perform a two-sample t test
e. Look at data using a histogram
*f. Perform a paired t test
g. Calculate a mean value
h. Calculate standard deviation

Calculate Question from *Chemical Engineering Thermodynamics*:

500 kJ of work is produced by expanding a system from 50 m3 to 60 m3 using a constant

pressure ____ kPa.

Correct Answer Range: 50 +/- 0

Fill-In-The-Blank Question from *General Chemistry*:

Write the names, with correct spellings, of the elements with the following symbols:

Ag _____ Au _____ Hg _____ Fe _____
Answers: silver, gold, mercury, iron

Fig. 1 Example assessment items

types varied across courses. Table 2 shows the distribution of question types. An example of each question type is shown in Fig. 1.

Three of the assessment items for each learning objective appeared on the practice quizzes (see below). The criterial test contained the fourth item of each of the 24 target learning objectives.

Procedure

There were numerous extra-experimental curricular differences between the courses involved in this research (e.g., different grading policies, numbers of exams, instructional approaches, etc.), but the assignment of items to the assessments (practice quizzes and criterial test) and the administration of the study materials did not vary.

Item assignment

Spacing was manipulated by assigning items to five biweekly quizzes in two different experimental conditions: *massed* and *spaced*. Half of the learning objectives were assigned to each condition. Learning objectives were numbered 1–24 in order of their appearance in the course, and all the odd-numbered objectives were assigned to one condition and all the even-numbered objectives to the other condition, in an alternating pattern. Assignment of objective to condition was counterbalanced across participants.

Table 3 Schedule of the experimental conditions: massed and spaced retrieval practice

Objectives	Condition	Quiz 1 (Week 3)	Quiz 2 (Week 5)	Quiz 3 (Week 7)	Quiz 4 (Week 9)	Quiz 5 (Week 11)	Criterial test (Week 14/15)
1–8	Massed	Questions 1, 2 & 3					Question 4
	Spaced	Question 1	Question 2	Question 3			Question 4
9–16	Massed		Questions 1, 2 & 3				Question 4
	Spaced		Question 1	Question 2	Question 3		Question 4
17–24	Massed			Questions 1, 2 & 3			Question 4
	Spaced			Question 1	Question 2	Question 3	Question 4

In the *massed* condition, all three practice items were assigned on the quiz immediately following the objective's introduction. In the *spaced* condition, the items were assigned to three consecutive quizzes, beginning with the quiz immediately following the objective's introduction. The spacing was therefore uniform, with a delay period of 2 weeks (see Latimier et al., 2021). Table 3 shows the temporal distribution of questions.

Administration of practice quizzes

At the beginning of the Fall 2020 semester, all students enrolled in each course were randomly assigned to one of two groups and assigned a set of practice quizzes. The five bi-weekly practice quizzes were administered after weeks 3, 5, 7, 9, and 11 via Blackboard®, the campus-wide learning management system. Quizzes were available to students from 1:00 p.m. on Friday afternoon until 11:59 p.m. Sunday night. Quizzes were not proctored, but several Blackboard® assignment settings were used to encourage retrieval of information as opposed to collaborative work or restudy (see Brothen & Wambach, 2004). Question order and multiple-choice answer order were randomized. Furthermore, once started, quizzes had to be completed in a fixed amount of time that was proportionate to the number of questions on the quiz (2 or 3 min per question for a 50- or 75-min class, respectively). Students had one attempt to complete each quiz. Lastly, feedback was not made available until 3:00 am the following Monday. Feedback consisted of the quiz questions, the possible answers, and whether each student's responses were correct or incorrect. For incorrect responses, the correct answer was not specified.

Throughout the semester, instructors sent out reminders to complete the study assignments. Members of the research team responded to any student issues that arose during the quiz-availability windows. In the case of a dropped wireless connection, a research team member recorded how many questions had been answered and reset the quiz to make it available again. Students were later removed from analysis if they had participated in additional retrieval practice opportunities due to connection problems and quiz resets.

Administration of the criterial test

The criterial test was given on the last day of class, which was either a Thursday or a Monday depending on the course schedule and was, respectively, 30 or 34 days after the fifth quiz window. The test was administered via Respondus Lockdown Browser®, a lockout proctoring system that blocks student use of the internet outside of the test. As on the practice quizzes, question order was randomized. When multiple-choice questions were asked, the multiple-choice answer order was also randomized. Backtracking to previously answered questions was prohibited. A member of the research team resolved and recorded any technical issues.

Partial credit was assigned for questions with multiple parts, with each complete question worth 1 point. For example, a question where a student got 2/3 parts correct would be an accuracy of 0.667. For multiple-choice items, including select-all-that-apply questions, a 0 or 1 was assigned.

Study-course integration

Instructors described the experimental materials in their syllabi as though they were regular elements of their courses. No distinction was drawn between the experimental components of the course and business-as-usual components. Note that the target learning objectives were also embedded into other components of the courses such as homework and periodic examinations.

In all courses, practice quizzes and the criterial test were valued, in total, between five and ten percent of students' final grades. The criterial test was not considered to be any course's final exam. Individual instructors chose

the exact value in this range. This value was intended to be sufficiently large to motivate students to complete the assignments. To further incentivize completion, all instructors offered a bonus to students who completed all practice quizzes and the criterial test. For example, one course valued the quizzes as 7% of the final grade and offered a bonus of 10% on the average (up to 100%) if students completed all assignments. Therefore, if a student's average on the quizzes was 75%, but the student completed all quizzes, the student's average was recorded as 85%, and this constituted 7% of their final grade in the class. Bonuses in other courses were similar.

At the end of the semester, in accordance with the IRB protocol, all students enrolled in the courses received an email explaining that they were automatically enrolled in a large research study that had no known risks.

Data analysis

Relevant performance data were exported from Blackboard® and demographic data were obtained from the university's Institutional Research Office. Raw data were deidentified and processed (see Fig. 2 for screening steps and analyses) using MySQL Workbench and R Studio (Boyd et al., 2021). Participants who experienced additional retrieval opportunities due to technical issues were removed from the dataset. The complete dataset is available on OSF (www.osf.io), and additional data summaries can be requested from the corresponding author.

Analysis 1. We first looked at student performance on the practice quizzes to determine whether spacing imposed difficulty during practice. For each student, separate averages were calculated for quiz questions in the massed and spaced conditions. For each course, these averages were submitted to a paired-samples *t* test. Hedges' *g* was our measure of effect size in these analyses. For this analysis, we used the complete set of practice data (3 questions for each of 24 learning objectives in each course) except for 2 learning objectives in the Unity of Life course that had to be removed due to errors during implementation. Higher accuracy on massed objectives indicated that spacing imposed difficulty during the practice phase.

Analysis 2. Next, we assessed the effect of spaced versus massed retrieval practice on criterial-test performance in each course. The dependent variable was proportion correct, calculated separately for test questions targeting objectives that received spaced practice (12 items) versus massed practice (12 items). Primary analyses consisted of eight paired-samples *t* tests, one for each course. Hedges' *g* was again the measure for effect size. Higher accuracy on spaced objectives indicated that spacing enhanced knowledge retention.

Analyses 3 and 4. Because our study was based on newly developed materials and was performed in real classrooms, we carefully reviewed the means, standard deviations, and skewness of responses to the questions on the criterial test. We looked especially for items that were at or near ceiling on accuracy, which we defined as above 0.90 proportion correct. High levels of accuracy are problematic for research purposes because they preclude the possibility of a spacing-induced increase in accuracy, but they are nonetheless a natural occurrence in real classrooms. Some learning objectives are likely to be mastered by large numbers of students, irrespective of any educational intervention. We flagged items that were at ceiling so we could conduct ancillary analyses with those items removed. In addition, we calculated the reliability of the 4 questions for each learning objective. We flagged learning objectives for which reliability was extremely low (Cronbach's $\alpha < 0.15$). Moreover, we examined whether the criterial-test question "hung together" with the quiz items by calculating reliability both with and without the criterial-test question. We flagged objectives for which inclusion of the criterial-test item was associated with a reduction in the Cronbach's α value of more than 0.10. We then reran the preliminary analyses (performance on the practice questions and criterial test in the massed and spaced conditions) with a filtered dataset that did not include any of the flagged items.

Meta-Analyses 1 and 2. After completing our primary analyses on courses individually, we assessed the generalizability of the effects using meta-analyses, first with the unfiltered data and then with the filtered data. Meta-analyses typically combine results (e.g., mean differences or effect sizes) from different articles that assess the same intervention to determine whether the effect is generalizable. A single-paper meta-analysis instead combines results from multiple studies within a single paper (McShane & Böckenholt, 2017). Either way, by considering multiple effects from different contexts, meta-analyses yield a more accurate estimate with decreased uncertainty and increased statistical power than the estimate of the effect of an individual study. To perform our single-paper meta-analyses, we submitted mean, standard deviation, and sample size data to the website: http://www.singlepapermetaanalysis.com/. These analyses tested whether spaced retrieval practice produced benefits in introductory STEM courses in general.

Meta-analyses 3 and 4. Lastly, we combined the data reported in this manuscript with data from our previous publication (Lyle et al., 2022) that reported the effectiveness of spaced retrieval practice in calculus. In the prior

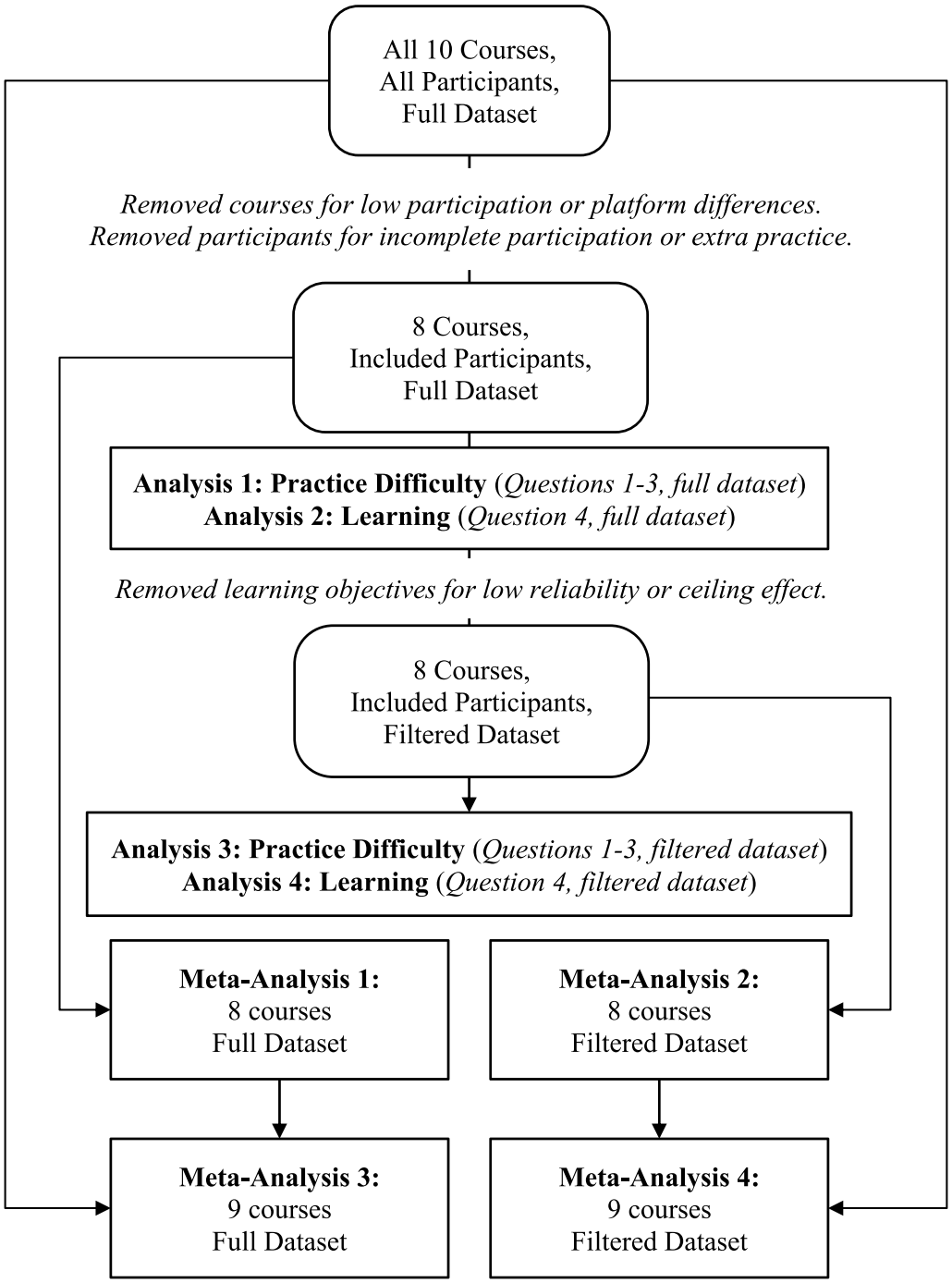

Fig. 2 Screening and analysis diagram

publication, spacing was administered through a different online platform, but otherwise had the same timing, number, and condition assignment of assessment items, the same study-course integration as described above, and was performed in the same semester (Fall 2020) at the same university. These analyses further investigate the generalizability of spaced retrieval practice in many introductory STEM classes.

Table 4 Practice quiz performance (percent correct) with results of statistical comparisons (full dataset)

Course	M (SD)		*t*	*p*	Hedges' *g*
	Massed	Spaced			
Chemical engineering thermodynamics	81.75 (10.94)	80.36 (12.2)	1.05	0.300	0.16
Chemistry for health professionals	89.77 (8.52)	88.77 (8.89)	0.92	0.360	0.09
Diversity of life	72.33 (13.95)	75.33 (12.77)	− 1.78	0.081	− 0.25
Fundamentals of physics I	76.73 (16.7)	76.81 (14.13)	− 0.08	0.939	− 0.01
General chemistry	86.8 (6.81)	86.93 (6.85)	− 0.17	0.868	− 0.02
Research methods for psychology	85.46 (9.13)	87.13 (9.47)	− 0.91	0.368	− 0.16
Statistics for psychology	79.43 (15.82)	76.32 (16.65)	1.85	0.068	0.21
Unity of life	83.63 (13.41)	79.98 (12.62)	3.53	< 0.001	0.35

The sign of the *t* and *g* values reflect "imposed difficulty", as calculated by massed performance minus spaced performance (i.e., *massed—spaced*). Thus, positive *t* and *g* values represent difficulty imposed by spacing, whereas negative values indicate that student performance was higher on spaced practice opportunities than massed practice opportunities

Table 5 Criterial test performance (percent correct) with results of statistical comparisons (full dataset)

Course	M (SD)		*t*	*p*	Hedges' *g*
	Massed	Spaced			
Chemical engineering thermodynamics	78.77 (13.30)	75.99 (15.75)	− 1.11	0.274	− 0.17
Chemistry for health professionals	82.67 (14.65)	85.77 (13.27)	1.81	0.073	0.17
Diversity of life	71.73 (15.28)	75.98 (17.21)	1.66	0.103	0.23
Fundamentals of physics I	77.75 (17.01)	76.10 (18.90)	− 1.21	0.229	− 0.12
General chemistry	81.11 (10.42)	84.64 (10.42)	1.98	0.052	0.25
Research methods for psychology	87.78 (14.96)	88.89 (13.37)	0.47	0.641	0.08
Statistics for psychology	70.95 (16.82)	72.80 (18.90)	0.93	0.357	0.11
Unity of life	78.76 (16.62)	78.79 (17.04)	− 0.34	0.735	− 0.03

The sign of the *t* and *g* values reflect "the learning gains due to spacing", as calculated by spaced performance minus massed performance (i.e., *spaced—massed*). Thus, positive *t* and *g* values represent higher performance on the criterial test due to spacing, whereas negative values indicate that student performance was higher after massed practice than after spaced practice

Results

The results sections are organized as follows: Analyses 1 and 3 addressed RQ1 (practice performance, with and without spacing), Analyses 2 and 4 addressed RQ2 (long-term knowledge retention, with and without spacing), and meta-analyses assessed the generalizability of the spacing effect across STEM courses.

Analysis 1: practice quiz performance—full dataset

A question of both theoretical and practical importance is whether student performance on practice opportunities was worse when these opportunities were spaced versus massed. Analysis revealed that this was not the case, at least not globally. As shown in Table 4, mean performance was lower in the spaced condition than the massed in only four courses and the difference was statistically significant in only one (Unity of Life). The effect size was small in all cases, including Unity of Life (also shown in Table 4). In some courses, performance was higher in the massed than the spaced condition, but in none of those cases was the difference statistically significant.

Analysis 2: criterial test performance—full dataset

Of primary interest was whether average student performance on the criterial test was better when the practice opportunities were spaced versus massed. Table 5 shows that this was the case in only five courses. The difference was not statistically significant in any of these. In the other three courses, mean performance was higher following massed quizzing but none of those differences were statistically significant.

Filtered data

Learning objectives were removed if performance on the fourth question was at ceiling, or if the set of four questions yielded low reliability (see the Method section for details). In total, 71 learning objectives were removed on these grounds (see Table 6).

Table 6 Number of items removed from the filtered dataset

Course	Q4 at ceiling	Low reliability	Total number removed
Chemical engineering thermodynamics	6	4	10
Chemistry for health professionals	11	1	12
Diversity of life	6	3	9
General chemistry	11	3	14
Fundamentals of physics I	2	0	2
Research methods for psychology	12	3	15
Statistics for psychology	5	1	6
Unity of life	4	1	5

Table 7 Mean criterial test performance (percent correct) with results of statistical comparisons (filtered dataset)

Course	M (SD)		*t*	*p*	Hedges' *g*
	Massed	Spaced			
Chemical engineering thermodynamics	70.46 (19.93)	67.62 (20.10)	− 0.86	0.394	− 0.13
Chemistry for health professionals	73.30 (21.01)	78.80 (19.50)	2.6	0.011	0.24
Diversity of life	65.25 (20.54)	71.13 (21.94)	1.69	0.097	0.23
Fundamentals of physics I	76.50 (17.83)	74.44 (19.92)	− 1.43	0.157	− 0.14
General chemistry	73.81 (15.17)	78.13 (16.48)	1.89	0.064	0.24
Research methods for psychology	75.95 (28.95)	79.64 (22.86)	0.77	0.451	0.14
Statistics for psychology	62.74 (21.17)	66.52 (23.29)	1.56	0.124	0.18
Unity of life	75.18 (20.95)	73.80 (20.96)	− 0.69	0.493	− 0.07

As in Table 5, the sign of the *t* and *g* values reflect "the learning gains due to spacing", as calculated by spaced performance minus massed performance (i.e., *spaced—massed*). Thus, positive *t* and *g* values represent higher performance on the criterial test due to spacing, whereas negative values indicate that student performance was higher after massed practice than after spaced practice

Analysis 3: practice performance—filtered dataset

Results were largely the same as those from the full dataset analysis. The only notable result was that the mean difference in Diversity of Life became significant, with performance in the spaced practice condition ($M = 75.67\%$) significantly higher than performance in the massed practice condition ($M = 70.92\%$), $t(50) = -2.40$, $p = 0.020$, *Hedges' g* $= -0.33$.

Analysis 4: criterial test performance—filtered dataset

We also reanalyzed criterial test performance using the filtered dataset. As shown in Table 7, mean performance in all courses was lower due to the removal of items that generated at-ceiling performance. The only difference in significance from the full dataset analyses was that student performance in Chemistry for Health Professionals was significantly better following spaced quizzing than massed.

Meta-analyses 1–4

To test the generalizability of spacing in STEM courses, we conducted single-paper meta-analyses of learning gains using the full (Meta-Analysis 1) and filtered (Meta-Analysis 2) datasets for the eight courses presented above. We then performed additional meta-analyses (Meta-Analyses 3 and 4) with the addition of the data from Calculus I for Engineering Students course presented by Lyle et al., 2022. The calculus data used in the analyses were as follows: $N = 180$; full data, $M_{spaced} = 77.05\%$, $SD_{spaced} = 17.45\%$, $M_{massed} = 71.44\%$, $SD_{massed} = 18.10\%$; filtered data, having removed 5 items at ceiling and 3 items with low reliability, $M_{spaced} = 70.78\%$, $SD_{spaced} = 21.64\%$; $M_{massed} = 64.35\%$, $SD_{massed} = 22.28$).

All mean values, confidence intervals, and results are presented in Fig. 3. The eight-course meta-analysis using the full dataset (Meta-Analysis 1) revealed a mean improvement of 1.50% (SE = 0.91%), with a 95% confidence interval of − 0.18 to 3.26%, which indicated that the effect was not significant. The heterogeneity value I2 of 87.7 (with an estimated range of 81.3–91.9%) indicated that 87.7% of the variance was between courses as opposed to between conditions. The results using the filtered data (Meta-Analysis 2) were similar; mean

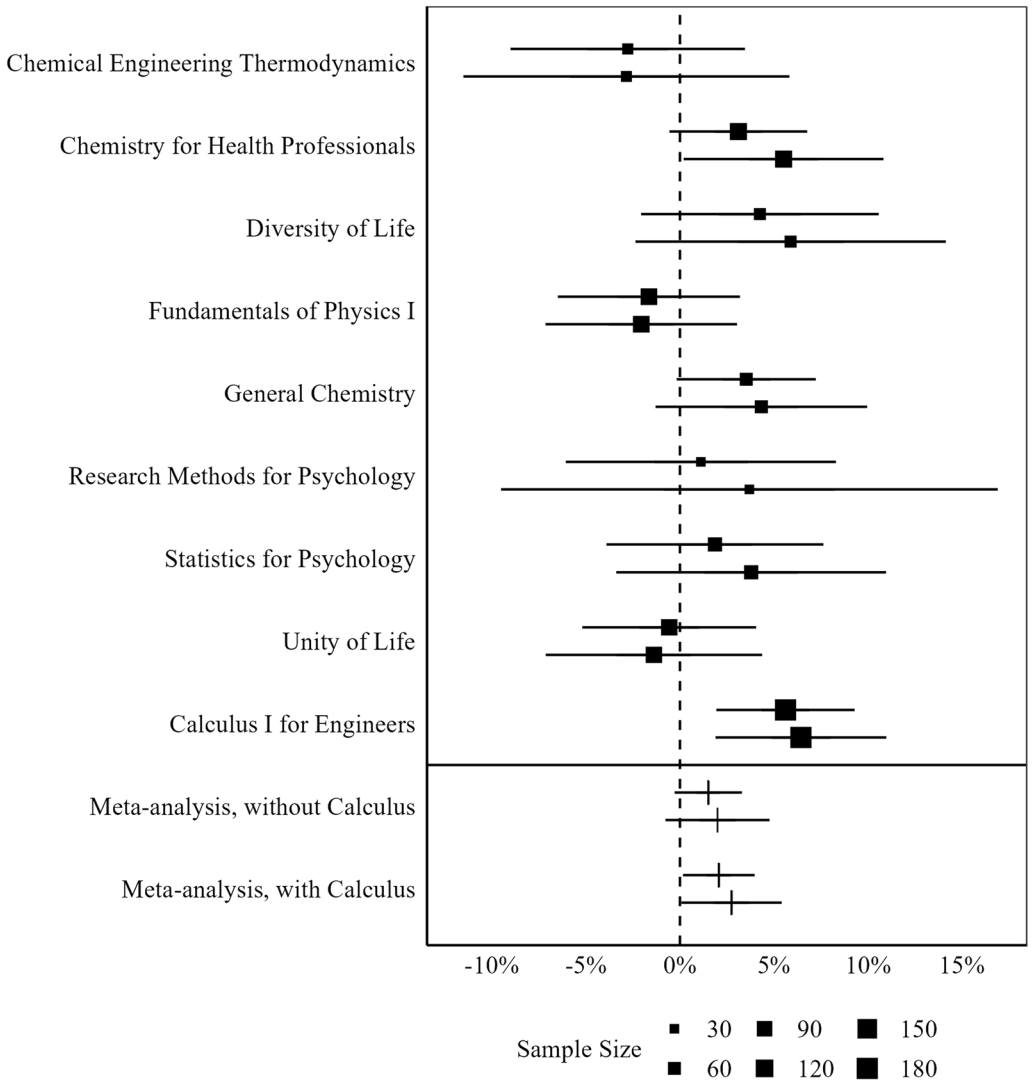

Fig. 3 Effect of spacing in nine STEM courses and meta-analytic results. Results from analyses of the full and filtered datasets are represented on the top and bottom lines, respectively, for each row

improvement = 2.00% (SE = 1.41%), 95% CI (− 0.77%, 4.76%), $I^2 = 74.68$.

The nine-course meta-analysis using the full dataset (Meta-Analysis 3) estimated the overall mean difference to be 2.06%, 95% CI (0.16%, 3.97%), indicating that spaced retrieval practice significantly increased student performance. However, the lower limit was not far above zero, and the heterogeneity value I2 was high at a value of 89.2% (95% CI: 84.3–92.6%), which showed that the results were highly variable across courses. The results using the filtered dataset (Meta-Analysis 4) were similar; mean improvement = 2.74% (SE = 1.36%), 95% CI (0.08%, 5.40%), $I^2 = 79.69\%$

Discussion

We sought to determine whether the memory-enhancing effect of spaced retrieval practice observed in the laboratory (Cepeda et al., 2006; Cull, 2000; Karpicke & Roediger, 2007; Landauer & Eldridge, 1967) and in mathematics classroom research (Hopkins et al., 2016; Lyle et al., 2020, 2022) would generalize across a variety of introductory STEM classrooms. We implemented spaced and massed retrieval practice on quizzes in nine different STEM courses using a within-subjects research design. We examined performance on an end-of-semester test as a function of whether prior quizzing of test content had been spaced or massed. We also examined performance

on the practice quizzes themselves, since it is often assumed that spacing makes retrieval more difficult during the practice phase of learning and thereby reduces practice performance (Bjork & Bjork, 2011; Bjork, 1994). In a previous publication (Lyle et al., 2022), we reported that participants in the calculus course performed significantly lower on practice quizzes but significantly higher on the criterial test in the spaced condition, indicating that spacing imposed desirable difficulty.

The results from the remaining eight courses of our study did not indicate consistent effects of spacing on either practice quizzes or the criterial test. Despite following the same spacing and massing schedule as in the calculus course, spacing significantly reduced practice performance in only one of the other eight courses (Unity of Life). When non-discriminative items were removed from analysis, spacing was shown to significantly improve practice performance in one course (Diversity of Life), and significantly increase criterial-test performance in one course (Chemistry for Health Professionals). No class yielded the pattern of results obtained in the calculus course (Lyle et al., 2022), even though the manipulation of massed versus spaced quizzing was the same in calculus as in the other courses.

The single-paper meta-analysis of the effect of spacing on criterial-test performance was not significant when considering results from the eight science and engineering courses presented in the current manuscript, of which only five had positive mean differences due to spacing. One could therefore conclude that the benefits of spacing as observed in the laboratory and in mathematics classrooms did not generalize across a variety of STEM subjects. On the other hand, the single-paper meta-analysis across all nine courses of the study (including calculus that was published previously; Lyle et al., 2022) revealed a significant and positive effect of spacing. In this case, 6 of 9 courses had positive effects, three of which had effect sizes greater than 0.20. One could therefore alternatively conclude that spacing *did* generalize across STEM disciplines. Ultimately, either of these conclusions about generalizability is an oversimplification. The proverbial glass is neither full nor empty, but rather somewhere in between. To fully capture the essential features and implications of our work, we consider generalizability of spacing from two perspectives.

The glass half-empty: the non-significant meta-analytic effect in eight courses

If one rejects the notion that the spacing effect generalizes across STEM courses, one is left to explain why robust effects were observed in mathematics classrooms (Emeny et al., 2021; Hopkins et al., 2016; Lyle et al., 2020, 2022) and not necessarily elsewhere. As alluded to in the introduction, the constituent STEM disciplines are grouped together not because they have shared cognitive underpinnings, but because they have collectively been prioritized by society (see McComas, 2014). But to say that subject matter (mathematics versus non-mathematics) is a moderator of the spacing effect in STEM classrooms is to beg the question of why discipline should matter. Is there something special about mathematics that makes it more sensitive than other subject areas to the temporal distribution of testing?

Based on data obtained in our research on calculus learning (Lyle et al., 2022), we previously proposed that spacing in mathematics benefits long-term retention by providing opportunities to integrate knowledge about older objectives with knowledge about newer objectives. We argued that this produced a more cohesive mental model and thereby benefitted learning (Soderstrom & Bjork, 2015). Perhaps such a process of model building, requiring the integration of older and newer knowledge, is especially critical for mastering complex, highly integrated bodies of mathematics knowledge, such as precalculus and calculus. We can foresee two responses to this idea: One, shouldn't we have obtained an effect in our statistics course, given the prominent role of mathematics in that course? Two, why have laboratory studies shown an effect of spaced retrieval in STEM subjects that are not centered around mathematics?

To the first question, about statistics, it is important to appreciate that much of the key content in statistics, especially at the beginning of the semester, which is when our intervention was applied, is conceptual, rather than computational (e.g., understanding the difference between descriptive and inferential statistics). In fact, only two of the 24 learning objectives chosen for study from the Statistics for Psychology course involved performing a calculation of any type.

To the second question, about laboratory studies, it is true that these have shown a spacing effect in multiple STEM subjects (e.g., Dobson et al., 2017; Grote, 1995; Reynolds & Glaser, 1964; Rohrer & Taylor, 2006, 2007). However, there may be critical differences between how students learn, or are taught, STEM topics in actual classrooms versus in the laboratory. The amount and scope of information to be learned in courses is much larger than in a laboratory study. In addition, how well students initially learn is not controlled in a classroom environment. This contrasts with a laboratory setup such as the one used by Rawson et al., (2013), where participants continue to study flash cards until they have established initial learning. Rawson et al. (2013) found a strong effect of spaced learning by repeatedly requiring students to demonstrate successful retrieval. Our study, like other classroom studies, did not require nor measure initial

learning. Therefore, students may not have attended to the information prior to answering retrieval practice items.

If initial learning is not established strongly in the classroom, students might need more support for learning outside of the retrieval practice opportunities such as immediate feedback to learn from the practice. Feedback has been shown to help students identify and correct their mistakes as well as gain confidence in correct responses (Butler et al., 2008; Pashler et al., 2005), which may be required in a classroom setting but not in laboratory studies (e.g., Roediger & Butler, 2011).

Other factors may also become important, such as question type (e.g., recall versus target recognition, Carpenter & DeLosh, 2006; Glover, 1989). Recently, Greving and Richter (2022) studied two question types in the context of a classroom research study where no feedback was given. They found positive results when using short-answer questions, and not when using multiple-choice questions. Therefore, they found evidence for a three-way boundary condition, in which retrieval practice did not benefit knowledge retention: (A) in a classroom setting, (B) when no feedback was provided after testing, and (C) when multiple-choice questions were the vehicle for retrieval.

Coincidentally, our current work fits squarely into this scenario. In most of the eight STEM classrooms, multiple-choice questions were the primary vehicle for retrieval (see Table 2) and, although feedback was made available to students, it was minimal, and students were not required to access it. This feedback structure may have prevented students from learning from spaced retrieval practice. In the calculus course, however, students were given more informative feedback, and questions were primarily fill-in-the-blank instead of multiple choice (see Lyle et al., 2022, for details). However, it is conceivable that the question type and feedback in the calculus course were better suited to promote learning than the spaced retrieval implementation in the other STEM courses. All cited spaced retrieval practice studies in mathematics classrooms made feedback available to students (e.g., Emeny et al., 2021; Hopkins et al., 2016; Lyle et al., 2020, 2022). The relative importance of feedback and question type and spacing in the classroom is unknown at this point, but these topics would be interesting for future research.

Also in a classroom context, there is the potential for participants to interact with the experimental materials on their own. One could posit that a spacing effect could be masked by pre-assessment studying. In addition, courses may have students with different characteristics that could introduce cross-course variation. For example, because medical school admission is strongly related to GPA, students in pre-medicine majors might be more inclined to study for small assignments (like the practice quizzes in our research) than engineering students, who might have higher workloads (see Lichtenstein et al., 2010) and whose career aspirations are not as strongly influenced by undergraduate GPA. In the cases where students were inclined and able to practice retrieval of target objectives outside of the practice quizzes, it is possible we would have been unable to detect a significant spacing effect. However, we have little ability to speak to this possibility because we did not collect relevant data on student study habits.

Therefore, our inconclusive results lead us to consider potential moderators of spaced retrieval practice, which could be subject matter, context, feedback, question type, or student study habits, among others. In addition, several moderators may interact to create boundary conditions for the effect of spaced retrieval practice. Although we would like to have investigated each of these already, our study design limits our ability to test for these moderators directly. We can only state that there appear to be limitations of the spacing effect across STEM courses, and we leave it for future research to reveal which factors are important and in what combination.

The glass half-full: the significant meta-analytic effect in nine courses

On the other hand, the meta-analysis of all nine courses in this study yielded a significant positive effect of spacing. There was a significant effect in Chemistry for Health Professionals, and positive, small-sized effects (> 0.20) in General Chemistry and Diversity of Life. Having a positive effect in these barrier courses could directly impact student enrollment and performance in upper-level STEM courses and also increase graduation rates in nursing, engineering, or pre-medicine disciplines.

These positive results were obtained in real college classrooms. Educational research in the classroom is messy when compared to a laboratory, with confounding and uncontrollable variables, distractions, and a larger scope of content. Students can decide when and how to interact with the course content outside of the carefully designed quizzing structure, and they may be externally motivated to do so to improve their grade. From a research perspective, it is therefore difficult to detect the impact of a real-world classroom intervention (see Taber, 2019). In addition, there are many differences between college courses besides subject matter (structure, instructor, and different samples of students) that may alter the effect of the spacing intervention. Significant results from classroom studies are therefore quite special, and our findings give credence to the practice of implementing spaced retrieval in the classroom.

In this case, we must again consider question type, and whether we could have found more or greater positive effects if we had used short-answer questions (Greving & Richter, 2022). This is possible. Therefore, instead of concluding that spaced retrieval practice either is or is not generalizable to all STEM courses, we prefer to state that more work is needed to determine when and how to apply spaced retrieval in the classroom for optimal benefit.

Limitations

Although this study tested spaced retrieval practice across many courses on a relatively large scale, it is limited to one experimental design. One concern is that something within our spacing manipulation, like interleaving the massed practice items within a quiz, may have masked the effect of spacing (Sana & Yan, 2022). Likewise, it is possible that we could have detected larger effects of spacing if we had included another, further-delayed assessment (Rohrer & Taylor, 2006), or a longer criterial test. Our study was subject to the practical constraints of experimentation within real-world classrooms such as the time available in class for the criterial test and the unlikelihood of continuing participation from students beyond the semester timeline. Another practical limitation of this real learning environment was that the learning objectives in the spaced condition may have been related to learning objectives in the massed condition. Practicing retrieval of information can increase retention of related, non-practiced information (Chan et al., 2006; Cranney et al., 2009; Rowland & DeLosh, 2014). However, if this relatedness, or any other course-necessitated element of this study reduced the observable impact of spacing, this is a limitation of the effectiveness of spaced retrieval practice and not of this study.

Our findings are limited, also, by the number of studies included, and the power available to detect significant differences within each course. According to a power analysis, a sample size of 27 participants is required to detect a significant effect using a paired t test with a power of 0.80, $\alpha = 0.05$, and medium effect size (Hedges' $g = 0.40$). Originally, we obtained ten high-enrollment courses (75+), but due to circumstances at the time, two of them did not reach this enrollment. One ultimately did not have enough participants to analyze (an undergraduate algebra course, $N = 11$), and another had a relatively low number of participants who completed all materials (Research Methods for Psychology, $N = 30$). The former was excluded, whereas the latter was included. The authors followed the a priori analysis plan with the intention of ethical research, specifically not running multiple analyses on the same dataset including and excluding different courses (to avoid p-hacking; Wicherts et al., 2016). Moreover, if removing one course changed the significance of either analysis, we would need to report all analyses, which would only add further to the "mixed results" message of this paper.

Lastly, this study was run in Fall 2020, the first semester following the outbreak of COVID. All courses were therefore conducted with remote elements, including remote proctoring of the criterial test. Students may have used cellphones or other resources to look up answers, reducing the evidence of a spaced retrieval practice effect. However, there was some degree of proctoring, and the value of the test was low. Moreover, remote test administration did not preclude obtaining a significant effect of spacing in calculus (Lyle et al., 2022) and Chemistry for Health Professionals. Consequently, we do not believe the possibility of cheating was a severe limitation of this research.

Conclusions and future directions

In this study, the generalizability of spacing-induced learning gains across STEM courses was unclear, but our results nonetheless raise important questions and may provide encouragement to some readers. Spaced retrieval practice is proving valuable in some STEM contexts. Existing research suggests the value of spacing in mathematics classrooms, and it showed small positive effects in three additional STEM courses in this study. What remains to be seen is whether potential learning gains of spacing are content-independent, as well as how best to design spaced retrieval manipulations in real learning environments. We have raised the possibility that the value of spacing may depend on the type of retrieval practice activity and feedback opportunities given to students, and future research should explore these implementation choices. In the meantime, we do not discourage instructors from experimenting (formally or informally) with spaced retrieval practice in STEM courses, but we caution them of the challenges in detecting positive returns.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-024-00468-5.

Additional file 1. Gender and racial demographics of the study sample, by course.

Additional file 2. Learning objectives for all courses.

Additional file 3. Assessment items for *Calculus I for Engineering Students*.

Additional file 4. Assessment items for *Chemical Engineering Thermodynamics*.

Additional file 5. Assessment items for *Chemistry for Health Professionals*.

Additional file 6. Assessment items for *Diversity of Life*.

Additional file 7. Assessment items for *Fundamentals of Physics I*.

Additional file 8. Assessment items for *General Chemistry*.

Additional file 9. Assessment items for *Research Methods for Psychology*.

Additional file 10. Assessment items for *Statistics for Psychology*.

Additional file 11. Assessment items for *Unity of Life*.

Acknowledgements
We would like to acknowledge the sequel (SQL) database support from undergraduate researchers Jeremy Boyd, Andrew Cleary, Osualdo Garcia, and Alvin Tran.

Author contributions
Authors include the principal investigators of the work as well as the instructor participants. This project would not have been possible without the hard work of all members of this team. KBL is now at TRANSFR VR, New York, NY.

Funding
The National Science Foundation: Improving Undergraduate STEM Education (NSF: IUSE) funded this work under award #1912253. Any opinions, findings, and conclusions or recommendations expressed in this manuscript are those of the PIs and do not necessarily reflect the views of the NSF; NSF has not approved or endorsed its content.

Availability of data and materials
All summative data and materials discussed herein are available through the Open Science foundation (OSF; https://osf.io/pkjf4/). Additional data are available upon request from the corresponding author.

Declarations

Competing interests
There were no competing interests or conflicts of interest in this work.

Author details
[1]Department of Engineering Fundamentals, University of Louisville, Louisville, KY, USA. [2]Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY, USA. [3]Department of Educational Leadership, Evaluation and Organizational Development, University of Louisville, Louisville, KY, USA. [4]Department of Physics and Astronomy, University of Louisville, Louisville, KY, USA. [5]Department of Chemistry, University of Louisville, Louisville, KY, USA. [6]Department of Biology, University of Louisville, Louisville, KY, USA. [7]Department of Mathematics, University of Louisville, Louisville, KY, USA. [8]Department of Chemical Engineering, University of Louisville, Louisville, KY, USA.

Received: 1 June 2023 Accepted: 16 January 2024
Published online: 07 February 2024

References

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

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review, 33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*(1), 43–56. https://doi.org/10.1037/xap0000133

Bacon, D. R., & Stewart, K. A. (2006). How fast do students forget what they learn in consumer behavior? A longitudinal study. *Journal of Marketing Education, 28*(3), 181–192. https://doi.org/10.1177/0273475306291463

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*. https://doi.org/10.1002/acp.3485

Bego, C. R., Ralston, P. A., Lyle, K. B., & Immekus, J. C. (2020). Research to practice to research: Intrinsic requirements of implementing and studying spaced retrieval practice in STEM courses. *October 2020 IEEE Frontiers in Education Conference (FIE)*, 1–5. https://doi.org/10.1109/FIE44824.2020.9273913

Bjork, E. L., Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*(59–68).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and Performance* (pp. 435–459). The MIT Press.

Boyd, J. R., Bego, C. R., Garcia, O., Ralston, P. A. S., Immekus, J. C., & Lyle, K. B. (2021). Using SQL to query the difficulty imposed by spaced retrieval in engineering mathematics. *IEEE Frontiers in Education Conference (FIE), 2021*, 1–4. https://doi.org/10.1109/FIE49875.2021.9637171

Brothen, T., & Wambach, C. (2004). The value of time limits on internet quizzes. *Teaching of Psychology, 31*(1), 62–64. https://doi.org/10.1207/s15328023top3101_12

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*(1), 69–79. https://doi.org/10.1007/s10734-010-9366-y

Burns, K. C., & Gurung, R. A. R. (2023). A longitudinal multisite study of the efficacy of retrieval and spaced practice in introductory psychology. *Scholarship of Teaching and Learning in Psychology, 9*(1), 96–103. https://doi.org/10.1037/stl0000206

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.34.4.918

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*(3), 369–378. https://doi.org/10.1007/s10648-012-9205-z

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition, 34*(2), 268–276. https://doi.org/10.3758/BF03193405

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *In Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: twelve years of cognitive psychology. *Journal of Experimental Psychology: General, 120*(4), 395–409. https://doi.org/10.1037/0096-3445.120.4.395

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*(6), 919–940. https://doi.org/10.1080/09541440802413505

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215–235. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3%3c215::AID-ACP640%3e3.0.CO;2-1

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1*(4), 309–330. https://doi.org/10.1007/BF01320097

Dobson, J. L., Perez, J., & Linderholm, T. (2017). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education, 10*(4), 339–347. https://doi.org/10.1002/ase.1668

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*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ebersbach, M., & Barzagar Nazari, K. (2020a). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *Journal of Applied*

Research in Memory and Cognition, 9(4), 532–541. https://doi.org/10.1016/j.jarmac.2020.08.014

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

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.

Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, M., Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, *4*(3), 1–24. https://doi.org/10.1177/25152459211027575

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399. https://doi.org/10.1037//0022-0663.81.3.392

Greving, S., & Richter, T. (2022). Practicing retrieval in university teaching: Short-answer questions are beneficial, whereas multiple-choice questions are not. *Journal of Cognitive Psychology, 34*(5), 657–674.

Grote, M. G. (1995). Distributed Versus Massed Practice in High School Physics. *School Science and Mathematics, 95*(2), 97–101. https://doi.org/10.1111/j.1949-8594.1995.tb15736.x

Gurung, R. A. R., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology, 33*(5), 732–743. https://doi.org/10.1002/acp.3507

Hartwig, M. K., Rohrer, D., & Dedrick, R. F. (2022). Scheduling math practice: Students' underappreciation of spacing and interleaving. *Journal of Experimental Psychology: Applied, 28*(1), 100–113. https://doi.org/10.1037/xap0000391

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

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin and Review, 12*(1), 159–164. https://doi.org/10.3758/BF03196362

Kamuche, F. U., & Ledman, R. E. (2005). Relationship of time and learning retention. *Journal of College Teaching & Learning (TLC), 2*(8), 25–28. https://doi.org/10.19030/tlc.v2i8.1851

Kang, S. H. K. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 12–19. https://doi.org/10.1177/2372732215624708

Karpicke, J. D., Butler, A. C., Roediger, H. L., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning Memory and Cognition, 33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Landauer, T. K., & Eldridge, L. (1967). Effect of tests without feedback and presentation-test interval in paired-associate learning. *Journal of Experimental Psychology, 75*(3), 290–298. https://doi.org/10.1037/h0025047

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

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Lichtenstein, G., Mccormick, A. C., Sheppard, S. D., & Puma, J. (2010). Comparing the undergraduate experience of engineers to all other majors: Significant differences are programmatic. *Journal of Engineering Education, 99*(4), 305–317. https://doi.org/10.1002/j.2168-9830.2010.tb01065.x

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

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

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

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

McComas, W. F. (2014). STEM: Science, Technology, Engineering, and Mathematics. In W. F. McComas (Ed.), *The Language of Science Education: An Expanded Glossary of Key Terms and Concepts in Science Teaching and Learning* (pp. 102–103). SensePublishers. https://doi.org/10.1007/978-94-6209-497-0_92

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*(1), 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*(6), 1048–1063. https://doi.org/10.1093/jcr/ucw085

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A. C., Koedinger, K. R., McDaniel, M. A., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning*. (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U. S. Department of Education. https://ies.ed.gov/ncer/pubs/practiceguides/20072004.asp

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning Memory and Cognition, 31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology, 55*(5), 297–308. https://doi.org/10.1037/h0040734

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

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*(4), 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 practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*(9), 1209–1224. https://doi.org/10.1002/acp.1266

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*(6), 481–498.

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

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin and Review, 21*, 1516–1523. https://doi.org/10.3758/s13423-014-0625-2

Sana, F., & Yan, V. X. (2022). Interleaving retrieval practice promotes science learning. *Psychological Science, 33*(5), 782–788. https://doi.org/10.1177/09567976211057507

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory and Cognition, 38*(2), 244–253. https://doi.org/10.3758/MC.38.2.244

Taber, K. S. (2019). Experimental research into teaching innovations: Responding to methodological and ethical challenges. *Studies in Science Education, 55*(1), 69–119. https://doi.org/10.1080/03057267.2019.1658058

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications, 3*(1), 1–17. https://doi.org/10.1186/s41235-017-0087-y

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology, 7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.