Introducing desirable difficulty in STEM barrier courses with spaced retrieval practice

Campbell R. Bego*
Department of Engineering Fundamentals
J. B. Speed School of Engineering
University of Louisville
campbell.bego@louisville.edu

Keith B. Lyle
Department of Psychological and Brain Sciences
College of Arts & Sciences
University of Louisville
keith.lyle@louisville.edu

Jason C. Immekus
Department of Educational Leadership, Evaluation and Organizational Development
College of Education
University of Louisville
jason.immekus@louisville.edu

Patricia A. S. Ralston
Department of Engineering Fundamentals
J. B. Speed School of Engineering
University of Louisville
patricia.ralston@louisville.edu

***Abstract*—This Full-Length Research Paper investigates the difficulty imposed by spaced retrieval practice in nine introductory Science, Technology, Engineering, and Mathematics (STEM) courses. By improving student performance in these courses, evidence-based pedagogical practices have the potential to increase graduation and success in STEM fields. *Spaced retrieval practice* is a technique in which questions on the same topic are asked repeatedly over time with intermittent delays. Spacing may initially make retrieval more difficult because it requires learners to recall information from long-term as opposed to short-term memory. However, this difficulty may ultimately be "desirable" because spacing often produces memory benefits in the long-term. The current paper examines the difficulty imposed by spaced retrieval in the nine STEM courses, using data collected from a 3-year project funded by the National Science Foundation. Results indicated that the magnitude of the difficulty imposed by spacing varied widely across the diverse STEM barrier courses. We anticipate that we will find similarly wide variability in the effectiveness of spaced retrieval practice in students' final learning outcomes, which will be investigated in future work.**

Keywords—spaced retrieval practice, distributed practice, spacing, desirable difficulty, STEM barrier courses

I. Introduction

Introductory Science, Technology, Engineering, and Mathematics (STEM) courses are intended to establish base knowledge and abilities for novices in these fields. Unfortunately, these fundamental courses are often difficult for students, and are thus referred to as "barriers" to STEM degrees (e.g., [1]). Due to anticipated STEM job growth in the US (e.g., [2]), instructional techniques that improve learning and success are needed to remove barriers to STEM degrees. Empirical research is needed to determine instructional techniques' effectiveness within and across STEM disciplines. A technique that is effective in one discipline may not be effective in another. This paper describes a large-scale study intended to assess the effectiveness of spaced retrieval practice in nine different STEM courses.

II. Literature Review

Spaced retrieval practice is an instructional technique in which students are asked to recall information related to the same topic multiple times with intervening temporal delays [3]. This technique capitalizes on two powerful findings from cognitive science: the *testing effect* and the *spacing effect*. The testing effect refers to the finding that the long-term retention of information is increased more by retrieving the information from memory than by restudying the information [4], [5]. Retrieval processes elicited by testing are the same as those required during a criterial test and, thus, can be considered transfer-appropriate [6]. Although the testing effect is well-established empirically, students rarely engage in self-testing on their own [7]. Consequently, it generally falls to instructors to provide opportunities for retrieval practice so that students can benefit from the testing effect. Classroom studies show that students retain more course content and earn higher grades when instructors require students to practice retrieval (e.g., [8]–[10]). Memory improves to an even greater extent when retrieval opportunities are spaced out over time versus *massed* in a short period of time [3], [11]. In the classroom, an example of massed practice is asking multiple questions about the same topic on a quiz. A meta-analysis [12] revealed a significant benefit of the spacing effect, even after taking publication bias into account. The average effect size was medium-to-large (*Hedges* $g = 0.74$).

Spacing requires recall from long-term instead of short-term memory. Retrieval from long-term memory can strengthen existing retrieval routes or create new ones, thereby improving future accessibility [13]. It can also make retrieval practice more difficult. Some researchers have observed lower performance on spaced retrieval attempts than massed ones [14]. Because spacing improves memory on criterial tests, lower performance

This project was funded by the National Science Foundation, Award # 1912253.

during practice might be considered *desirable difficulty* [15].The premise of desirable difficulty is that more difficult exercises during the early stages of learning can improve long-term knowledge retention [16], [17]. Evidence-based techniques that increase the initial difficulty of learning and result in improved long-term retention include: varying the conditions of practice, reducing feedback, interleaving content, and spacing [18].

Unfortunately, like testing, spacing is not often implemented by learners, instructors, or the designers of instructional materials. For example, a recent review of mathematics textbooks revealed predominantly "chunked" layouts, with each end-of-chapter problem set related only to local chapter content [19]. It is therefore again up to instructors to discover and implement spacing for themselves.

A. *Spaced retrieval practice in STEM courses*

Spaced retrieval practice has been well established in the laboratory, especially in the domain of verbal learning [3]. Only recently have researchers begun to study spacing implementations in STEM classrooms (e.g., [20]–[23]). Unlike participants in the laboratory, students in STEM courses have a great amount of control over their learning, making it more difficult to detect any effect of a classroom manipulation. However, evidence-based instructional techniques are meant to assist real-world learning. Therefore, educational psychology research in the classroom context is necessary to validate laboratory results, and findings have important implications for instructors and researchers alike [24].

One such applied study examined the effectiveness of two different evidence-based practices in an undergraduate engineering mathematics course [22]. In that study, the number of quiz questions targeting various learning objectives was manipulated, constituting a manipulation of amount of retrieval practice. Also manipulated was the temporal distribution of the questions such that multiple questions targeting the same learning objective were presented on a single quiz or on multiple quizzes spaced across several weeks. This constituted a manipulation of the spacing of retrieval practice. Analysis of performance on an end-of-semester test revealed that mastery of learning objectives was greater when students had answered more (versus fewer) questions during the semester and when those questions were spaced across multiple quizzes (versus massed on a single quiz). The effect of spacing was greater than the effect of increasing the amount of practice and was longer lasting. On a test administered at the beginning of the following semester, mastery was found to be greater for learning objectives that had been the subject of spaced versus massed practice. Amount of prior practice was unrelated to mastery. Additional analyses revealed that performance on spaced quiz questions was lower than performance on massed quiz questions. Greater difficulty during practice coupled with superior performance on a long-term test supports the idea that spacing imposed desirable difficulty during the initial learning phase [22].

To-date, no studies have investigated the effectiveness of spaced retrieval practice across multiple STEM domains. Consequently, pragmatic research in STEM is limited and, in general, cross-disciplinary work is rare. The current study is designed to fill this gap.

B. *The Current Study*

We implemented spaced retrieval practice in 9 introductory STEM courses. The courses included foundational courses in engineering, mathematics, physics, biology, chemistry, and psychology. In the current paper, we explore whether spacing imposes difficulty on practice quizzes in these courses. Our research questions were:

Does spacing reduce performance during retrieval practice opportunities in these 9 STEM barrier courses?

Does the magnitude of spacing-induced difficulty vary by course or discipline?

We hypothesized that spacing would lower performance on retrieval practice opportunities in some courses. In a prior ASEE poster, we assessed the spacing-induced difficulty in a single course and found it to be significant [25]. We expected variability in the magnitude of spacing-induced difficulty across courses but did not formulate a priori hypotheses about the magnitude in specific courses.

It is important to note that we did not assess long-term learning in this paper, and therefore we are not able to say whether difficulty is desirable. However, these intermediate results are important for our understanding of the mechanism by which spaced retrieval practice may enhance learning.

III. Methodology

A. *Participants*

Participants (*N* = 790) were students who (a) were enrolled in one of the nine STEM barrier courses at the University of Louisville in Fall 2020, (b) completed all five practice quizzes, and (c) had no computer or internet access errors that affected the manipulation. One additional course was excluded from the analysis because of a low number of participants (*N* = 11). The number of participants and basic demographic statistics (% male and % white) in each course are reported in Table 1.

TABLE 1. DEMOGRAPHIC STATISTICS

Course	**Course Name**	***N***	**% Male**	**% White**
Biology 1	Unity of Life	108	35%	65%
Biology 2	Diversity of Life	56	25%	57%
Chemistry 1	Chemistry for Health Professionals	120	11%	67%
Chemistry 2	General Chemistry	62	39%	74%
Engineering Thermodynamics	Chemical Engineering Thermodynamics	43	58%	88%
Engineering Mathematics	Calculus I with Engineering Analysis	183	73%	77%
Physics	Fundamentals of Physics I	110	40%	65%
Psychology 1	Quantitative Methods in Psychology	76	28%	63%
Psychology 2	Research Methods for Psychology	32	44%	75%
Total		790	41%	70%

B. Procedures

In the first year of this three-year project, instructors attended workshops to learn about spaced retrieval practice and how to select or develop test questions for classroom administration. The research team helped instructors select target learning objectives from the first half of the semester, and assisted in the creation or selection of associated assessment questions. Each instructor identified 24 target learning objectives (eight that would be introduced during weeks one to three, four to five, and six to seven) and wrote or selected four questions for each learning objective. For challenges experienced and lessons learned during this development process, please see [26].

Once questions were finalized, quizzes were created to implement the spacing manipulation. The manipulation was within-subjects, such that each student experienced 12 of the target learning objectives in a *spaced condition*, and 12 in a *massed condition* (Table 2). In the massed condition, three of the four questions about the learning objective were asked on the quiz immediately following content instruction. In the spaced condition, the three questions were administered in three consecutive quizzes beginning with the quiz immediately following content instruction. In both conditions, the fourth question was asked on the final, criterial assessment of learning. Only performance on the first three questions of each objective are considered in this paper.

Two sets of quizzes were created to counterbalance objectives across conditions. At the beginning of the Fall 2020 semester, students were randomly assigned to receive one set of questions or the other, resulting in the creation of two groups arbitrarily labeled A and B. Groups A and B differed only in which objectives were assigned to the massed or spaced condition.

Practice quizzes were administered from Friday 1:00 pm to Sunday 11:59 pm, following weeks three, five, seven, nine, and eleven of the semester.

C. Materials

Questions. As stated above, course instructors developed or selected all learning objectives and test questions for this study. Question types were either multiple choice or fill in the blank. To receive a list of these materials, please contact the corresponding author.

Quizzes. Five practice quizzes were administered using the Blackboard© Learning Management System for all but one course, which instead utilized Pearson's MyLabsPlus®. Quizzes consisted only of questions covering the target learning objectives in the massed and spaced conditions, as described in Table 2. The number of questions varied by quiz, as follows: Quiz 1, 16 questions; Quiz 2, 20 questions; Quiz 3, 24 questions; Quiz 4, eight questions; and Quiz 5, four questions. The order of questions on each quiz was randomized for each student. Multiple-choice answers were also randomized. Students were given a limited amount of time to complete quizzes after opening them. Time was proportional to the number of questions asked and the length of the class (50, 75, or 110 minutes). Quizzes were not proctored.

D. Data Analysis

We compared student performance on practice questions in the massed and spaced conditions. We first calculated an average score for each condition (12 learning objectives * 3 practice questions = 36 items) for each student. Difficulty was operationalized as the difference in student performance on practice questions between the spaced and massed conditions (*spaced – massed*; a larger positive difference represents greater difficulty). We assessed the difficulty imposed by spacing in each course with a paired-samples t test that compared the condition averages. Cohen's d was used as a measure of effect size.

Prior to running the comparative analyses, we checked for violations to the assumptions of a paried t test. The research design was such that the dependent variables were continuous and the observations were independent. Outliers were considered using Z scores of the difference scores, calculated independently for each course. Only one Z score was more than 3 standard deviations from the mean, with a value of 3.04. As this value was not extreme, and as there were several datapoints immediately below 3 standard deviations from the mean, it was not considered an outlier and was not removed from the dataset.

The difference scores in each class were normally distributed (Shapiro-Wilk, $p > .05$), except for Biology 1, Shapiro-Wilk

TABLE 2. SPACING MANIPULATION

Condition	**Content**	**Time administered**					
		Quiz 1	**Quiz 2**	**Quiz 3**	**Quiz 4**	**Quiz 5**	**Quiz 6**
Massed	Half of the LOs from weeks 1-3	Question 1 Question 2 Question 3					Question 4
	Half of the LOs from weeks 4-5		Question 1 Question 2 Question 3				Question 4
	Half of the LOs from weeks 6-7			Question 1 Question 2 Question 3			Question 4
Spaced	Half of the LOs from weeks 1-3	Question 1	Question 2	Question 3			Question 4
	Half of the LOs from weeks 4-5		Question 1	Question 2	Question 3		Question 4
	Half of the LOs from weeks 6-7			Question 1	Question 2	Question 3	Question 4

Note: This table illustrates how questions were assigned in a series of five practice quizzes and a final criterial quiz. In the massed condition, three questions were assigned on the same quiz. In the spaced condition, three questions were assigned over three consecutive quizzes. Each student received half of the 24 target learning objectives (12, four per unit) in the massed condition, and half in the spaced condition. LO stands for Learning Objective.

$W(108) = .973$, $p = .029$. Because *t* tests are robust to normality vioations with larger sample sizes (see [27], [28]), we expected that a parametric analysis would adequately assess significance and also produce a meaningful difference score, thereby providing ecological validity. However, we also ran a follow-up nonparametric related-samples Wilcoxon Signed Rank test for this course. Results are reported below.

IV. Results & Discussion

In this preliminary investigation, we assessed the difficulty imposed by spacing varied across 9 STEM barrier courses. Results from paired-sample *t* test statistics are presented in Table 3. As anticipated, there was variability in the difficulty associated with spacing, from -2.88% (the negative value indicates that student performance in the spaced condition was higher than the massed condition) to 3.23%. The mean difficulty imposed by spacing was .74% (SD = 2.04).

In Biology 1 and Engineering Mathematics, the difficulty imposed by spacing was statistically significant (see Table 3). The follow-up nonparametric Related-Samples Wilcoxon Signed Rank test of Biology 1 data also indicated a significant impact of spacing, $Z(277.3) = 3190.5$, $p = .003$. Spacing-related performance reductions in these courses were 3.19% and 2.69%, respectively. The spacing-related performance reduction in Psychology 1 was of similar magnitude (3.23%) but was not significant. On the other end of the spectrum was Biology 2, in which students performed better in the spaced condition than the massed condition (-2.88%), but not significantly. Results from the other courses showed little, if any, difficulty imposed by spacing.

The difficulty observed in the Engineering Mathematics course replicates other work in engineering mathematics [22]. Obtaining this effect in a different mathematics course is an important step towards generalization within the mathematics domain. It is even more promising because the referenced study also found statistically significant improvement in learning due to spacing. Therefore, spacing may provide desirable difficulty in engineering mathematics. This is an exciting finding due to the critical nature of mathematics in engineering. Despite historical attempts to make calculus a “pump” instead of a “filter” [29], early undergraduate mathematics courses are often barriers to success in engineering [30], [31]. Our results indicate that spaced retrieval practice may be a useful tool in increasing student success in engineering.

Variability across courses in the difficulty associated with spacing should motivate exploration of differences between courses that may cause differences in the effects of spacing. Differences in implementation can be ruled out because all courses utilized the same number of objectives, questions, and quizzes, the same timing, and the same within-subjects manipulation. In addition, quizzes in all courses were assigned the same weight towards students’ overall grade (5-10%). However, many differences between courses remain. One possibility is that spacing effects differ across STEM domains. Our study was intended to compare across domains, and we are finding differences. However, there is no theory to suggest that spacing would create difficulty in one domain over another, so we do not believe that domain is the source of the differences observed. Other more justifiable possibilities are (1) sample variations between courses (age, major, race, gender) or (2) differences in the amount of intrinsic spacing within each course.

Samples differ across the STEM courses in this study. Although all courses were selected as introductory STEM courses, they differed in their placement along curricular paths, with some expected to be taken earlier and others later. Thus, some courses primarily consist of sophomore and junior students, whereas others have primarily freshman enrollment. Spacing may have a stronger effect for first-year students who may not have established their study strategies yet, especially in the very first semester of college (fall). For instance, older students may have developed more effective study strategies that minimize the effectiveness of the spaced retrieval practice implementation. It is not likely that older students have acquired spaced retrieval as a study strategy [32], however, they may have learned to pay close attention to feedback on any assignments. If older students had in fact learned to pay attention to available feedback, they may perform better on later, spaced questions. This idea is supported by the difficulty observed in Engineering Mathematics, where enrollment consisted primarily of first-year

TABLE 3. RESULTS

Course	**N**	**Mean Difference**[a]	**SD**	**95% CI**		***t***	***df***	***p***	***Cohen’s d***	**95% CI**	
				Lower	**Upper**					**Lower**	**Upper**
Biology 1	108	3.19%	10.33	1.22	5.16	3.21	107	**.002**	.31	0.11	0.50
Biology 2	56	-2.88%	11.86	-6.05	0.30	-1.82	55	.075	-.24	-0.51	0.02
Chemistry 1	120	0.29%	11.96	-1.87	2.45	0.26	119	.792	.02	-0.15	0.20
Chemistry 2	62	-0.06%	6.06	-1.60	1.48	-0.08	61	.939	-.01	-0.26	0.24
Engineering Thermodynamics	43	1.03%	8.53	-1.59	3.66	0.79	42	.431	.12	-0.18	0.42
Engineering Mathematics	183	2.69%	12.10	0.92	4.45	3.01	182	**.003**	.22	0.08	0.37
Physics	110	0.18%	10.60	-1.83	2.18	0.17	109	.862	.02	-0.17	0.20
Psychology 1	76	3.23%	14.37	-0.05	6.52	1.96	75	.053	.23	0.00	0.45
Psychology 2	32	- 1.04%	10.00	-4.65	2.56	-0.59	31	.560	-.10	-0.45	0.24

[a] The mean difference reported here represents the difficulty imposed by spacing, The difficulty score was calculated by subtracting the average performance on practice questions in the spaced condition (36 items) from the average performance on practice questions in the massed condition (36 items) by student, i.e., *massed – spaced,* and then taking the average of the difference for all students.

students. In contrast, Engineering Thermodynamics primarily consisted of sophomores. We do not have student age or year information for the other courses, so we cannot yet determine the impact of student age on the difficulty of spaced retrieval practice. This is a viable direction for future work, however.

Student major may similarly relate to the magnitude of the difficulty of spacing. Students who desire to go to medical school, for example, know that they must perform well in their courses to generate competitive applications. Even a 5% difference in final grade can be the difference between acceptance and rejection in the medical field. Thus, these students may go out of their way to perform better on intermediate, low-stakes assignments. Several of the smaller difficulty scores were seen in courses required for pre-med majors (Biology 2, Chemistry 1, Chemistry 2, and Physics). However, Biology 1, which showed a statistically significant imposed difficulty of spacing, is also required for medical school applications. The impact of major on the effectiveness of spaced retrieval practice is of interest due to the number of jobs anticipated in various fields. According to the Bureau of Labor Statistics [33], job growth in the next ten years (2019-2029) is predicted to be extremely high in Health Services, the highest of any category of professions. Therefore, helping learners perform better in introductory science courses required for the medical field can assist in filling these job openings.

Additional sample differences that may affect the impact of spacing are demographics such as gender or race. As seen in Table 1, courses had different enrollment by both demographic factors. To the authors' knowledge, there are no studies that directly look at the impact of spaced retrieval practice across demographics. However, recent work in mathematics showed a gender difference in learning through testing [34]. The study looked at whether a test provided a learning opportunity by comparing student performance on two consecutive tests, varying the amount of content-related questions in the first test. Gains in performance from content-related testing were larger for males than females. It is possible that we will observe gender differences in the spaced retrieval practice manipulation. We plan to perform follow-up analyses to look for gender differences in the difficulty imposed by spacing.

Outside of sample differences, a second major possible explanation is the amount of intrinsic spacing within courses. The Physics course, for example, began with 4 weeks of instruction in vector math, which was then utilized for the rest of the semester. It would not be possible to solve problems later in the semester without using information from the beginning of the semester. Therefore, spaced retrieval is naturally embedded in Physics. It is likely that intrinsic spacing would reduce the difficulty of a spaced retrieval practice manipulation. The Engineering Thermodynamics instructor also believed the course to have a large amount of intrinsic spacing, but so too did the Engineering Mathematics instructor. To investigate this further, it will be important to develop an objective measure of intrinsic spacing. We would need to determine whether and how much problems later in the semester require retrieval of information presented earlier in the semester.

A third possible explanation lies in different question formatting across courses. Some instructors wrote or selected multiple choice questions, whereas others utilized fill-in-the-blank and multi-select questions. Multiple-choice questions have a built-in floor for accuracy, with the level of chance at 25%. Fill-in-the-blank alternatively allows accuracy to drop to zero. This discrepancy between question formats may impact our ability to discern difficulty at the small amounts of 3-5%. Also with respect to the questions asked across courses, intrinsically easier questions (of any format) could have showed less spacing-induced difficulty. It is possible that easier questions, when spaced, remain easy to answer over time. Further investigation of our data to detect differences in item difficulty are necessary to explore this potential moderator of the difficulty of spaced retrieval practice.

Overall, the results from this study indicate that there may be moderators of the difficulty imposed by spacing across STEM domains. Further investigation of these factors will help illuminate the theory behind the spacing effect and help instructors determine whether or not they wish to implement spacing.

A. Limitations

One limitation of this study is that it does not look at the level of difficulty of the courses, learning objectives, or items. It could be that some quiz questions were very easy such that most or all students answered them correctly regardless of whether they were asked in spaced or massed fashion. The presence of any such items would reduce our ability to detect spacing-induced difficulty. Item analysis will be conducted in the future to identify potentially problematic questions.

It is also important to remember that we have not yet assessed the final outcomes of the intervention. We do not yet know whether spacing improved performance in the long run, only that it imposed variable amounts of difficulty in the short run across STEM barrier courses. In future work, we will assess whether difficulty correlated with learning gains in the final assessment.

V. Conclusions & Future Work

This paper presented an exploration of the difficulty imposed by spaced retrieval practice in introductory STEM courses. We found differences across courses, which indicate that there may be moderators of the impact of spacing in various courses. We described some of these potential moderating factors, and conclude that further analyses are needed to determine predictors of the difficulty imposed by spacing.

Planned future work includes investigation of student age, gender, and major, as well as whether the spacing manipulation impacted learning in the final assessment.

Acknowledgments

We would like to acknowledge the hard work of the instructors throughout this project: Raymond Chastain, Lora Haynes, Lenore Hoyt, Rachel Pigg, Shira Rabin, Matthew Scobee, and Thomas Starr. We would also like to thank the undergraduate research support team that helped with implementation, data collection, and data organization: Clayton Robertson, Haasya Kanamarlapudi, Osvaldo Garcia, and Jeremy Boyd. Lastly, we appreciate the suggestions of the anonymous FIE reviewers.

REFERENCES

[1] A. Redmond-Sanogo, J. Angle, and E. Davis, "Kinks in the STEM Pipeline: Tracking STEM Graduation Rates Using Science and Mathematics Performance," *Sch. Sci. Math.*, vol. 116, no. 7, pp. 378–388, 2016, doi: 10.1111/ssm.12195.

[2] R. Noonan, "STEM jobs: 2017 Update. ESA issue brief # 02-17.," *US Dep. Commer.*, 2017.

[3] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," *Psychol. Bull.*, vol. 132, no. 3, pp. 354–380, 2006, doi: 10.1037/0033-2909.132.3.354.

[4] C. A. Rowland, "The effect of testing versus restudy on retention: A meta-analytic review of the testing effect.," *Psychol. Bull.*, vol. 140, no. 6, pp. 1432–1463, 2014, doi: 10.1037/a0037559.

[5] K. B. McDermott, "Practicing retrieval facilitates learning," *Annu. Rev. Psychol.*, vol. 72, no. 1, pp. 609–633, 2021, doi: 10.1146/annurev-psych-010419-051019.

[6] C. D. Morris, J. D. Bransford, and J. J. Franks, "Levels of processing versus transfer appropriate processing," *J. Verbal Learning Verbal Behav.*, vol. 16, no. 5, pp. 519–533, 1977.

[7] J. Dunlosky and R. Ariel, "Self-regulated learning and the allocation of study time," *Psychol. Learn. Motiv.*, vol. 54, pp. 103–140, 2011.

[8] F. C. Leeming, "The exam-a-day procedure improves performance in psychology classes," *Teach. Psychol.*, vol. 29, no. 3, pp. 210–212, 2002, doi: 10.1207/S15328023TOP2903_06.

[9] K. B. Lyle and N. A. Crawford, "Retrieving essential material at the end of lectures improves performance on statistics exams," *Teach. Psychol.*, vol. 38, no. 2, pp. 94–97, 2011, doi: 10.1177/0098628311401587.

[10] M. A. McDaniel, J. L. Anderson, M. H. Derbish, and N. Morrisette, "Testing the testing effect in the classroom," *Eur. J. Cogn. Psychol.*, vol. 19, no. 4–5, pp. 494–513, 2007, doi: 10.1080/09541440701326154.

[11] N. Kornell, "Optimising learning using flashcards: Spacing is more effective than cramming," *Appl. Cogn. Psychol.*, vol. 23, no. 9, pp. 1297–1317, 2009, doi: 10.1002/acp.1537.

[12] A. Latimier, H. Peyre, and F. Ramus, "A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention," *Educ. Psychol. Rev.*, pp. 1–29, 2020, doi: 10.1007/s10648-020-09572-8.

[13] S. K. Carpenter, "Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval.," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 35, no. 6, p. 1563, 2009.

[14] J. D. Karpicke and H. L. Roediger, "Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 33, no. 4, pp. 704–719, 2007, doi: 10.1037/0278-7393.33.4.704.

[15] G. B. Maddox and D. A. Balota, "Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty," *Mem. Cogn.*, vol. 43, pp. 760–774, 2015, doi: 10.3758/s13421-014-0499-6.

[16] E. L. Bjork and R. A. Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," in *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2nd Editio., M. A. Gernbacher and J. Pomerantz, Eds. New York, NY: Worth, 2014, pp. 59–68.

[17] R. A. Bjork, "Memory and metamemory considerations in the training of human beings.," in *Metacognition: Knowing about knowing.*, J. Metcalfe and A. Shimamura, Eds. Cambridge, MA: MIT Press, 1994, pp. 185–205.

[18] R. A. Bjork, "Assessing our own competence: heuristics and illusions," in *Attention and performance XVII: Cognitive regulation of performance:Interaction of theory and application*, D. Gopher and A. Koriat, Eds. Cambridge, MA: The MIT Press, 1999, pp. 435–459.

[19] D. Rohrer, R. F. Dedrick, and M. K. Hartwig, "The scarcity of interleaved practice in mathematics textbooks," *Educ. Psychol. Rev.*, vol. 2020, pp. 1–11, 2020.

[20] R. F. Hopkins, K. B. Lyle, J. L. Hieb, and P. A. S. Ralston, "Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge," *Educ. Psychol. Rev.*, vol. 28, no. 4, pp. 853–873, 2016, doi: 10.1007/s10648-015-9349-8.

[21] R. A. R. Gurung and K. Burns, "Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice," *Appl. Cogn. Psychol.*, vol. 33, no. 5, pp. 732–743, 2019, doi: 10.1002/acp.3507.

[22] K. B. Lyle, C. R. Bego, R. F. Hopkins, P. A. S. Ralston, and J. L. Hieb, "How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge," *Educ. Psychol. Rev.*, vol. 32, no. 1, pp. 277–295, 2019, doi: 10.1007/s10648-019-09489-x.

[23] A. C. Butler, E. J. Marsh, J. P. Slavinsky, and R. G. Baraniuk, "Integrating cognitive science and technology improves learning in a STEM classroom," *Educ. Psychol. Rev.*, vol. 26, no. 2, pp. 331–340, 2014, doi: 10.1007/s10648-014-9256-4.

[24] R. E. Mayer, "What good is educational psychology? The case of cognition and instruction," *Educ. Psychol.*, vol. 36, no. 2, pp. 83–88, 2001.

[25] C. R. Bego, P. A. S. Ralston, K. B. Lyle, and J. C. Immekus, "Introducing desirable difficulty in engineering mathematics with spaced retrieval practice," in *Proceedings of the 128th ASEE Annual Conference and Exposition*, 2021, pp. 1–7.

[26] C. R. Bego, P. A. Ralston, K. B. Lyle, and J. C. Immekus, "Research to practice to research: Intrinsic requirements of implementing and studying spaced retrieval practice in STEM courses," in *October 2020 IEEE Frontiers in Education Conference (FIE)*, 2020, pp. 1–5.

[27] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *Int. J. Endocrinol. Metab.*, vol. 10, no. 2, p. 486, 2012.

[28] J. Pallant, *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge, 2020.

[29] L. A. Steen, Ed., *Calculus for a new century: a pump, not a filter. Papers from a national colloquium (MAA Notes No. 8, October 28-29, 1987)*. Washington, DC: Mathematical Association of America, 1988.

[30] J. Moore, "Undergraduate mathematics achievement in the emerging ethnic engineers programme," *Int. J. Math. Educ. Sci. Technol.*, 2005, doi: 10.1080/00207390500063975.

[31] P. K. Subramaniam, M. Cates, and G. Borislava, "Improving success rates in calculus," *MAA Focus*, vol. 28, no. 5, pp. 20–21, 2008.

[32] R. N. Blasiman, J. Dunlosky, and K. A. Rawson, "The what, how much, and when of study strategies: Comparing intended versus actual study behaviour," *Memory*, vol. 25, no. 6, pp. 784–792, 2017.

[33] B. of L. Statistics, "News release: Employment projections—2019-2029," Washington, DC: US Department of Labor. [USDL-20-1646], 2020.

[34] F. Avvisati and F. Borgonovi, "Learning mathematics problem solving through test practice: A randomized field experiment on a global scale," *Educ. Psychol. Rev.*, vol. 32, no. 3, pp. 791–814, 2020.