# Can Interleaving Enhance Student Outcomes in Freshman Calculus?

Steve Bennoun
University of California Los Angeles

*In contrast to traditional homework sets that typically group problems by topics, an approach known as blocking, interleaving consists in alternating the topics for each problem set. Studies in middle school mathematics and college physics report important benefits of interleaving on students outcomes. The goal of the present study is to investigate whether similar benefits can be observed in the context of college calculus. Students in two sections of a Calculus for Life Sciences course received either blocked or interleaved weekly homework. On two surprise quizzes given in the middle and at the end of the quarter, the interleaved group outperformed the blocked group (d=0.18 and d=0.33 respectively). On a quiz given one month after the end of the course, the interleaved group outperformed the blocked group on the second half of the quarter but not on the first half. Limitations of these results are discussed.*

*Keywords:* student achievement, calculus, interleaving effect.

The United States faces an urgent need to improve mathematics and science education in order to meet the future demand of science, technology, engineering and mathematics (STEM) workforce (President's Council of Advisors on Science and Technology (PCAST), 2012; National Research Council, 2013). Over the years, there has been many calls for the development and adoption of evidence-based instructional methods in STEM courses, especially at the college level (e.g., National Research Council, 2000; American Association for the Advancement of Science, 2009). Poor performance in introductory math courses such as calculus has been linked to students' decisions to leave STEM majors (Chen, 2013; Ehrenberg, 2010; Hunter, 2019). Calculus is often considered a "weed out" course (Weston et al., 2019) and constitutes one of the most challenging hurdles for students pursuing STEM majors. Improving student outcomes in calculus is therefore of great importance.

There is strong evidence that active learning can improve student learning compared to traditional methods (e.g., Freeman et al., 2014; Code et al., 2014; Hake, 1998). However, active learning does not by itself guarantee that all students will automatically perform better or earn top grades. It is therefore necessary to investigate other teaching methods that can be used either by themselves or in combination with other instructional approaches such as active learning.

One component of instruction that is often overlooked by teachers and students is how to organize and sequence learning opportunities. A common way of organizing a course is to divide topics into blocks that are then studied and practiced individually, an approach known as blocking (Rohrer, Dedrick, & Hartwig, 2020). In contrast, interleaving consists in mixing the practice of different skills together. Thus, when completing an interleaved homework set, students practice both the individual skills in the current section and also the ability to identify which method (from across prior units) is appropriate to solve a given problem. This is not trivial because similar looking problems may require different strategies (e.g., $\int x(e+1)^x dx$ requires integration by parts whereas $\int e(x+1)^e dx$ requires a U-substitution) and choosing an appropriate strategy is often challenging for students (Ziegler & Stern, 2014). In addition, exams are typically interleaved as they cover a range of different topics and students are not explicitly told which concept or method they should use to solve a given problem. Students are therefore expected to exhibit on exams a skill that has hardly (if at all) been practiced. There is a growing body of literature from cognitive science and education psychology showing the potential of interleaving to support student learning and performance. The goal of this study is to investigate how interleaving can be concretely implemented in a college Calculus for Life Sciences course and what its impact on student performance is.

## Theoretical Background and Relevant Literature

The first study comparing interleaving and blocking for mathematical tasks was conducted by Rohrer and Taylor (2007). Participants in this study learned to compute the volumes of four little-known solids following either a blocked or interleaved schedule. In the blocked group, the participants read a tutorial on how to compute the volume of a given solid and then completed four practice problems on that solid. They repeated this pattern for the four solids. In contrast, in the interleaved group participants first read the four tutorials and then completed the practice problems that were interleaved, or in other words, mixed. In both groups participants received feedback after each problem in the form of the correct answer. The participants took part in two practice sessions that were one week apart. During these sessions the blocked group showed greater accuracy with an average of 89% versus 60% for the interleaved group. However, on the test that took place one week after the second practice session, the scores dramatically reversed with the blocked group having an average score of 20% and the interleaved group 63%.

Two non-mutually exclusive mechanisms have been proposed to explain the benefits of interleaving: the *spacing effect hypothesis* and the *discriminative contrast hypothesis*. Since interleaving consists in intermixing the practice of different topics, an interleaved schedule automatically creates some temporal spacing between two practice problems or practice sessions on a given concept. Many studies have shown the benefit of spacing repetitions over time rather than massing repetitions (e.g., Cepeda et al., 2006; Donovan & Radosevich, 1999; Bjork, 1994). The *discriminative contrasts hypothesis* proposes that juxtaposing problems on different concepts next to each other promotes the comparison between the concepts and makes their discriminative features more salient. Several studies support this hypothesis (e.g., Wahlheim et al., 2011; Sana et al., 2017) and two in particular (Taylor & Rohrer, 2010; Kang & Pashler, 2012) suggest that the benefits of interleaving do not solely come from the spacing effect.

Large-scale studies conducted in the classroom with authentic classroom materials have reported a large positive effect of interleaving. For example, Rohrer et al. (2020) conducted a study with 54 seventh-grade classes involving 787 students. The students practiced four different types of mathematical problems during 8 practice sessions. Each class either followed a blocked or interleaved practice schedule. Students then completed a test 33 days after the last practice session. On that test, the interleaved group significantly outperformed the blocked group with an average of 61% versus 38% (Cohen's d=0.83). While showing a large effect, this study focused only on four topics and was conducted in middle school. One question is thus whether a similar positive effect of interleaving can be observed in a course where all topics throughout the course are interleaved and where the material is more complex.

Samani and Pan (2021) addressed this question by interleaving or blocking homework sets in an undergraduate physics course. During four weeks, students would either complete blocked or interleaved homework sets. Then for the next four weeks they would switch to other the type of homework. Students received a surprise test two days after the end of each stage. On both tests the interleaved group outperformed the blocked group (Cohen's d=0.40 for the first stage, and d=0.91 for the second stage). One should note all problems were computational in nature with a numerical final answer (with units) and involved recalling the appropriate formula. One should also note that no statistically significant differences were observed on the course exams. In this study students completed three homework sets per week for a total of 27 problems, which is more than the typical one homework set per week, and over 75% of them were seniors.

The goal of the present exploratory study is to test whether interleaving could bring similar benefits in an introductory Calculus for Life Sciences course in which students are mainly freshmen and sophomores and complete one homework set per week. Moreover, questions should also include non-computational questions such as writing the equations of a mathematical model based on the description of a biological system or interpreting a bifurcation diagram. More specifically, I answer the following two research questions:

- What is the impact of interleaving on student performance in a Calculus for Life Sciences course in which students complete one weekly homework?
- What are the benefits of interleaving homework sets after a one-month delay?

## Methods

This study was conducted in two sections of a Calculus for Life Sciences course at a large research-intensive public university in the southwest with a total of 491 students. The vast majority of the students (94%) were either in their freshman or sophomore year and over 90% planned to major in biology or the life sciences. The course focused on modeling and how to use mathematical models to better understand biological systems. The instructor used active learning methods in the form of clicker questions with peer instruction (Crouch & Mazur, 2001) and small-group activities. The students also had weekly coding labs led by teaching assistants. During these labs students solved one question similar to a homework exercise in groups, and then spent rest of the time learning to code in SageMath. The course was identical for both sections (same instructor, same lecture materials, etc.) except for the structure of the homework. The ten-week quarter was divided into two four-week stages (the topics of weeks 5 and 10 were not part of the study). During the first stage, one section received blocked homework sets while the other one received interleaved homework sets. Each section then switched to the other type of homework for the second stage.

Blocked and interleaved homework sets were created in the following way. Each lecture was divided into two topics and two or three homework exercises were created on each topic. In the blocked homework sets the topics appeared in the same order as they had been studied in class and the exercises on a given topic were grouped together. For the interleaved homework sets, once a topic was studied in class, one exercise on this topic would appear on the next three homework sets (except for the last two weeks of each stage). The interleaved homework sets thus contained topics from different weeks and two exercises on a given topic would never appear one after the other. Importantly, in each stage students received the exact same homework exercises, only organized differently. Students submitted their homework weekly and homework exercises were graded on a 3-point scale (very good, sufficient, insufficient).

Students received a surprise in-class quiz on the Thursday of weeks 5 and 10. They also had a midterm on the Monday of week 6 (just before the beginning of the second stage) and a final exam on the Monday following week 10. In order to assess the impact of interleaving after a one-month delay, students took an in-class quiz during the first day of the following quarter. This quiz took place 28 days after the final exam. Students were encouraged to do their best on that quiz by being told that it would help the instructors assess their current level of understanding of the material was. They also received a bonus point for completing the quiz. The quiz contained two questions on material from the first stage and two questions on material from the second stage. The midterm, final exam and quizzes were graded by the teaching assistants of the course.

## Results

### Quizzes and Exams During the Quarter

For the first surprise quiz, students in the interleaved group outperformed the students in the blocked group with an average of 66.11% for the interleaved group and 62.14% for the blocked group ($p=0.047$, Cohen's $d=0.18$). A similar outcome was observed on the second surprise quiz that took place on the last day of the course with an average of 57.24% for the interleaved group and 50.54% for the blocked group ($p<0.001$). This time the effect size was larger with Cohen's $d=0.33$. One should note that while the second half builds on the first half, the questions asked in each quiz targeted concepts from the corresponding stage. A student can do well on a stage 2 question (e.g., determining the long-term behavior of a given model using linear stability analysis) without necessarily doing well on a stage 1 question (e.g., writing down the equations of a model based on the description of a biological system or estimate the future value of a variable using Euler's method). It thus seems reasonable to assume that the results of the second quiz are primarily (if not only) impacted by the interleaving or blocked condition in stage 2.

For the midterm and final exam, there was no statistically significant differences between the two groups. Specifically, on the midterm the interleaved group had an average of 79.81% and the blocked group of 79.42% ($p=0.78$). For the final exam, restricting the analysis to the questions covering the second half of the quarter, students in the interleaved group outperformed students in the blocked group with an average of 87.68% for the interleaved group and 85.11% for the blocked group. However, this difference was not significant ($p=0.092$). The difference between the two groups on the questions covering the first part of the quarter was not significant either.

Although the lack of statistically significant differences between the two groups on the midterm and final exam may suggest a limitation of the impact of interleaving, two other elements may explain this result. First, it is possible that the unannounced quizzes helped students identify topics they had yet to master and thus prompted them to thoroughly review these concepts. This hypothesis is consistent with results on topics that appeared both on quizzes and the subsequent exam. For example, for the question on the stability of equilibrium points the average was 43.57% on the end-of-quarter quiz and 89.14% on the final exam. Second, even without the unannounced quizzes, students tend to review very intensely shortly before the exam. On surveys students indicated having spent on average 16.63 and 21.63 hours specifically reviewing for the midterm and the final exam respectively. One should also note that there is no indication that study behavior changed between the two groups. For example, the perceived difficulty of the homework reported by the students was not significantly different between the two groups. This suggests that the differences seen on the quizzes does not come from different study behaviors but from interleaved structure of the homework itself.

### Quiz After a One-Month Delay

The quiz that was given at the beginning of the following quarter contained four questions, two on material from the first stage and two on material from the second stage. On the first stage there was no statistically significant difference between the two groups ($p=0.79$).

On the second stage the interleaved group outperformed the blocked group with an average of 56.71% for the interleaved group and 51.75% for the blocked group ($p=0.044$, Cohen's $d=0.20$). There is some similarity with the results of the quizzes given during the quarter with interleaving having a stronger impact on the second half of the quarter than on the first half. This is also in line with the results reported by Samani and Pan (2021) in college physics.

## Discussion and Future Work

The results suggest that interleaving can help student performance in calculus (although perhaps modestly) and are consistent with previous studies in middle-school mathematics (Rohrer, Dedrick, Hartwig, et al., 2020; Rohrer et al., 2015) and college physics (Samani & Pan, 2021). These three studies have reported larger gains and one question is therefore where this difference may come from. The first study by Rohrer et al (2015) focused only on two types of problems, graph and slope problems. One interesting feature is that the wording of each type of problem was always the same ("Graph the equation…" and "Find the slope of the line that passes through the points…"). It is reasonable to assume that the problems given on the test again used the same phrasing. While not explicitly mentioned, it seems probable that this same feature was used by Rohrer and collaborators in their (2020) study. This characteristic leaves open the possibility that interleaving may help students better identify keywords or key features of problems but is not necessarily as powerful when problems on a given topic can take different formats. Based on the examples accompanying their article, Samani and Pan (2021) also seem to have used problems that had the exact same wording except for the numbers and names of the atoms involved in the problems. Since their tests involved some problems that combined two topics, it is hard to determine whether the wording of the test problems was similar to the homework problems. However, since in the present study some problems on a given topic could have different format, this could explain the lower effect sizes observed here.

Another notable feature of Samani and Pan's (2021) study is that all problems involved recalling the correct formula(s) and then using it to get to the correct numerical value and units. When looking at their memory sub-measure (whether students could recall the appropriate formula), they found effect sizes of $d=0.41$ for the first stage and $d=0.96$ for the second stage, while when looking at the correctness of the answers sub-measure, the effect sizes were $d=0.25$ for the first stage and $d=0.40$ for the second stage. While the authors explain that it is not uncommon in introductory physics courses to have a low level of correct numerical answers due to the conceptual and computational complexity of the questions, the difference in effect sizes may suggest that interleaving really helps students recall the appropriate formulas but is not as potent in supporting their problem-solving skills.

One common feature between the current study and Samani and Pan (2021) is that the interleaving effect is stronger for the second half of the course. There is no obvious explanation for this phenomenon, however, one may tentatively hypothesize that interleaving helps students more when the complexity of the material increases. This is something that needs to be further tested. Another question that needs further investigation where do the differences in effect sizes between the current study and the studies by Rohrer et al. (2020; 2015) and Samani and Pan (2021) come from. Is it due to the differences in the format of the problems used as noted above or do other factors also play a role? Finally, future research should further look into why benefits observed on unannounced quizzes do not carry over to exams and if this is indeed due to students cramming before an exam or due to other characteristics of the exams.

## Acknowledgements

This work is supported by the National Science Foundation (NSF) grant DUE-2225258. All findings and opinions are those of the author and not necessarily of the NSF.

## References

American Association for the Advancement of Science. (2009). *Vision and change in undergraduate biology education: A call to action*. https://web.archive.org/web/20230121040940/https://visionandchange.org/wp-content/uploads/2013/11/aaas-VISchange-web1113.pdf

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing.* (1994-97967-009; pp. 185–205). The MIT Press. https://doi.org/10.7551/mitpress/4561.003.0011

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. http://dx.doi.org/10.1037/0033-2909.132.3.354

Chen, X. (2013). *STEM Attrition: College Students' Paths Into and Out of STEM Fields* (NCES 2014–001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED544470

Code, W., Piccolo, C., Kohler, D., & MacLean, M. (2014). Teaching methods comparison in a large calculus class. *ZDM-Mathematics Education*, *46*(4), 589–601. https://doi.org/10.1007/s11858-014-0582-2

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977. https://doi.org/10.1119/1.1374249

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Ehrenberg, R. G. (2010). Analyzing the factors that influence persistence rates in STEM field, majors: Introduction to the symposium. *Economics of Education Review*, *29*(6), 888–891. https://doi.org/10.1016/j.econedurev.2010.06.012

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Hunter, A.-B. (2019). Why Undergraduates Leave STEM Majors: Changes Over the Last Two Decades. In E. Seymour & A.-B. Hunter (Eds.), *Talking about leaving revisited: persistence, relocation, and loss in undergraduate STEM education* (pp. 87–114). Springer International Publishing. https://doi.org/10.1007/978-3-030-25304-2_3

Kang, S. H. K., & Pashler, H. (2012). Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. The National Academies Press. https://www.nap.edu/catalog/9853/how-people-learn-brain-mind-experience-and-school-expanded-edition

National Research Council. (2013). *The Mathematical Sciences in 2025*. The National Academies Press. https://doi.org/10.17226/15269

President's Council of Advisors on Science and Technology (PCAST). (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology,* *engineering, and mathematics*. Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review*, *32*(3), 873–883. https://doi.org/10.1007/s10648-020-09516-2

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40–52. https://doi.org/10.1037/edu0000367

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900–908. https://doi.org/10.1037/edu0000001

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. https://doi.org/10.1007/s11251-007-9015-8

Samani, J., & Pan, S. C. (2021). Interleaved practice enhances memory and problem-solving ability in undergraduate physics. *Npj Science of Learning*, *6*(1), 1–11. https://doi.org/10.1038/s41539-021-00110-x

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*(1), 84–98. https://doi.org/10.1037/edu0000119

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848. https://doi.org/10.1002/acp.1598

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750–763. https://doi.org/10.3758/s13421-010-0063-y

Weston, T. J., Seymour, E., Koch, A. K., & Drake, B. M. (2019). Weed-Out Classes and Their Consequences. In E. Seymour & A.-B. Hunter (Eds.), *Talking about leaving revisited: Persistence, relocation, and loss in undergraduate STEM education* (pp. 197–243). Springer International Publishing. https://doi.org/10.1007/978-3-030-25304-2_7

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, *33*, 131–146. https://doi.org/10.1016/j.learninstruc.2014.04.006