Picking the Right Dose: The Challenges of Applying Spaced Testing to Education

Douglas P. Larsen, MD, MEd

Two of the most rigorously and extensively studied learning principles in cognitive psychology are the effects of spacing and testing.[1–3] The spacing effect refers to the principle that when studying or testing is distributed over time, information is more likely to be retained than if studying or testing is massed over a short period of time.[3] Studies have also shown that when information is recalled from memory during testing, retention is better than if the material is simply studied (ie, the testing effect). The effort required to retrieve information from memory is thought to make the memory trace more enduring (for this reason the term *retrieval practice* is sometimes used to describe the testing effect).[1] Multiple laboratory studies as well as studies in educational settings have shown that spacing and testing can be used in combination to increase long-term retention.[4–6] These tools have been shown to be effective for a wide range of learners: older adults,[7] middle school students,[8] medical students,[9] and residents.[6] In addition, spacing and testing have been shown to be effective whether the same questions are used as a final assessment or novel tests are used.[6,9,10] In short, the effects of spacing and testing are quite robust because they apply to many types of learners under many different situations.

Despite the consistent benefits demonstrated in most studies, the application of these principles can be nuanced. Finding the optimal spacing interval can present challenges. If tests are too widely spaced, then considerable forgetting occurs and students must subsequently relearn the material. If tests are not spaced enough, then testing will not require sufficient mental effort to produce the desired effects on memory. For instance, flash cards used back-to-back may lead to short-term improvements in retention after a week but have no effect when retention is measured 6 months later.[11] Because the flash cards are used in relatively close succession, there is little spacing between retrieval attempts. Studies using tests spaced by 1 or more weeks have shown improved retention over 6 months or more.[6,9,10] Typically, to achieve retention over months or years, testing must be spaced by weeks or even months.[5] However, as previously mentioned, educators may need to start with more frequently spaced tests to ensure initial learning, and then expand to more widely spaced tests to solidify retention.

The other question that educators must consider is how many repetitions are necessary to retain information over the long term. Studies that have directly manipulated the number of testing repetitions have found that long-term retention substantially increased the more often items were retrieved from memory.[4] As educators consider using spaced testing, they may find that more difficult materials require more repetition. Thus, to use testing as an effective learning tool, educators must identify the number of testing repetitions that are appropriate for their materials and practical for their circumstances.

Close examination of the data from several studies demonstrates that not all topics benefit equally from spaced testing in the research setting. In some studies of spaced testing in which multiple topics were used, an interaction between topics and interventions was found.[9,10] Despite the fact that learning improved for most topics when spaced testing was used, at least 1 topic in each of these studies showed equivalent performance between groups that had spaced testing and control groups that repeatedly studied the material. No groups performed worse with testing than studying. The likely reason for this variation in effects is due to the nature of the materials for each topic. Unlike laboratory studies, which use tightly controlled materials, such as word lists or brief passages of text, the studies referenced here used actual materials from medical education. Inherently there was more variability in the nature of the materials. Some topics required lists of medications and others required extensive knowledge of physical examination techniques. Research studies—whether in the laboratory or with actual educational materials—by necessity use the same protocol for each topic and condition. However, in real life, topics cover a wide range of difficulty. Difficult topics may require different strategies, such as more frequent testing or different schedules in spacing. Educators must adapt and modify research findings to fit their own needs and content.

This brings us to the study by Mathes et al[12] in this issue of the *Journal of Graduate Medical Education*. The authors found that spaced testing with multiple-choice questions led to different effects between 2 courses covering different

Douglas P. Larsen, MD, MEd, is Assistant Professor, Division of Pediatric and Developmental Neurology, Department of Neurology, Washington University School of Medicine.

Corresponding author: Douglas P. Larsen, MD, MEd, Washington University School of Medicine, 660 South Euclid, Campus Box 8111, St. Louis, MO 63110, 314.454.6120, larsend@neuro.wustl.edu

DOI: http://dx.doi.org/10.4300/JGME-D-14-00170.1

topics in pediatric dermatology. In 1 course, spaced testing seemed to result in improved knowledge, with a 15% higher score in the intervention group producing a moderate effect size (d = 0.45) and very nearly reaching statistical significance (P = .06). These results suggest that the spaced testing most likely had an impact, but that the study was not sufficiently powered to reach statistical significance. This inference is strengthened by the fact that the study included an inadequate number of participants to meet the numbers derived by the authors' a priori power calculation. Similar results in other studies have provided evidence of the long-term effects of testing.[13] In the authors' second course, they found that those who had received the course scored worse than those who had no exposure to the information; this difference was also not statistically significant. This divergence in results is consistent with a topic effect in the second course, where the materials and learning strategy may not have been the right fit to produce a gain in knowledge. As with the studies mentioned above, the course material may have required more repetitions or an alternative spacing of tests in order to produce the long-term effect in learning that the course designers intended.

The study by Mathes et al[12] demonstrates that one size does not fit all topics in applying spaced testing. These results and others raise a couple of important questions for educators: (1) When and how should spaced testing be used in educational programs? and (2) If spaced testing does not produce a substantial difference in all instances, should it be used at all? First, educators must ensure that they are using spaced testing optimally based on the existing evidence. This evidence suggests that for tests to be used as learning tools they must be aligned with educational objectives, allow for adequate repetition, when possible use questions that require leaners to generate answers rather than recognize information, provide feedback, and ensure appropriate spacing.[14] When spaced testing is optimized, the empirical evidence suggests that spaced testing leads to at least equivalent performance and in most cases superior performance compared with other learning interventions. Educators may have to try various regimens of spaced testing to find the best fit for their materials. They may even find that some materials are best learned through entirely different means. At the end of the day, spaced testing does not replace good educational planning and instructional design. Educators must still ensure that they are choosing the right "medicine" and the right "dose" to achieve their objectives.

References

1. Roediger HL III, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci*. 2006;1:181–210.
2. Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. *Med Educ*. 2008;42(10):959–966.
3. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol Bull*. 2006;132(3):354–380.
4. Rawson KA, Dunlosky J. Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *J Exp Psychol Gen*. 2011;140(3):283–302.
5. Cepeda NJ, Vul E, Rohrer D, Wixted JT, Pashler H. Spacing effect in learning: a temporal ridgeline of optimal retention. *Psychol Sci*. 2008;19(11):1095–1102.
6. Larsen DP, Butler AC, Roediger HL III. Repeated testing improves long-term retention relative to repeated study: a randomised, controlled trial. *Med Educ*. 2009;43(12):1174–1181.
7. Balota DA, Duchek JM, Paullin R. Age-related differences in the impact of spacing, lag, and retention interval. *Psychol Aging*. 1989;4(1):3–9.
8. McDaniel MA, Agarwal PK, Huelser BJ, McDermott KB, Roediger HL III. Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *J Educ Psychol*. 2011;103:399–414.
9. Larsen DP, Butler AC, Lawson AL, Roediger HL III. The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ Theory Pract*. 2013;18(3):409–425.
10. Larsen DP, Butler AC, Roediger HL III. Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ*. 2013;47(7):674–682.
11. Schmidmaier R, Ebersbach R, Schiller M, Hege I, Holzer M, Fischer MR. Using electronic flashcards to promote learning in medical students: retesting versus restudying. *Med Educ*. 2011;45(11):1101–1110.
12. Mathes EF, Frieden IJ, Cho CS, Boscardin CK. Randomized controlled trial of spaced education for pediatric residency education. *J Grad Med Educ*. 2014;6(2):270–274.
13. Kromann CB, Bohnstedt C, Jensen ML, Ringsted C. The testing effect on skills might last 6 months. *Adv Health Sci Educ Theory Pract*. 2010;15(3):395–401.
14. Larsen DP, Butler AC. Test-enhanced learning. In: Walsh K, ed. *Oxford Textbook of Medical Education*. Oxford, UK: Oxford University Press; 2013:443–452.