Effect of selected “desirable difficulty” learning strategies on the retention of physiology information

John L. Dobson
Department of Health and Kinesiology, Georgia Southern University, Statesboro, Georgia

Submitted 5 May 2011; accepted in final form 25 July 2011

Dobson JL. Effect of selected “desirable difficulty” learning strategies on the retention of physiology information. Adv Physiol Educ 35: 378–383, 2011; doi:10.1152/advan.00039.2011.—The purpose of this study was to examine the effects of interleaving and expanding retrieval on the retention of physiology concepts. Participants (n = 189) read and then reread 30 immunology and reproductive physiology passages. Half of the participants read and then reread the passages in a blocked manner (e.g., a1a2a3b1b2b3), and the other half did so in an interleaved manner (e.g., a1b1a2b2a3b3). Participants were then repeatedly assessed, without feedback, after either a uniform or an expanding series of intervals. Half of the students from both the blocked and interleaved groups completed the assessments 1, 2, and 3 days after rereading the passages (uniform), whereas the other half completed the assessments immediately and 1 and 3 days after rereading the passages (expanding). All participants completed a final assessment 10 days after rereading the passages. There were no significant differences between the blocked and interleaved groups on any of the assessments, nor were there any significant interactions between the groups on any of the assessments. Those in the expanding retrieval group scored significantly higher than those in the uniform group on all four assessments (ANOVA; assessment 1: F = 35.12, P = 0.00; assessment 2: F = 13.88, P = 0.00; assessment 3: F = 10.87, P = 0.00; and assessment 4: F = 6.79, P = 0.01). Mean final assessment scores were 47.58 ± 19.81 and 40.50 ± 17.17 for the expanding and uniform groups, respectively. The results indicate that participants benefited more from expanding retrieval practice.

interleaving; expanding retrieval

SOME OF THE MOST ROBUST FINDINGS in the field of experimental psychology indicate that learning is enhanced when learners are prompted to more actively process the information they are internalizing. This effect is perhaps best demonstrated with so-called “desirable difficulties,” which are learning strategies that have been advocated by experimental psychologist Dr. Robert Bjork. These strategies require learners to exert additional effort when acquiring or retrieving information, which may actually slow initial learning, but the evidence indicates they tend to result in greater long-term gains in learning. Two intriguing, but relatively controvertible, examples of desirable difficulties are interleaving and expanding retrieval practice.

The practice of interleaving means to structure the acquisition of multiple learning sets such that learning alternates between sets. In contrast, it is far more typical to structure learning sets in a blocked manner, such as A1, A2, A3, A4, B1, B2, B3, B4. With interleaving, those same sets of information might be structured in a progression such as A1, B1, A2, B2, A3, B3, A4, B4. Much of the early empirical support for learning from interleaved versus blocked practice comes from studies that investigated motor skills. For example, studies have demonstrated that interleaving results in superior learning associated with general motor skills (18), with baseball batting skills (8), and with the retention and transfer of badminton skills (7). More recent evidence has also indicated that interleaved practice can result in an enhanced ability both to solve math problems (15, 20) and recognize unique artistic styles in paintings (11). However, the effect of interleaving on the retention of words and language and the ability to learn scientific concepts is less clear. To the best of the author’s knowledge, only two pertinent investigations have been published thus far, and those two studies had inconsistent results.

That is, the first of those two studies (3) found that interleaving resulted in superior recall of information (i.e., word pairs), but the subsequent study by Richland et al. (14) found that the practice did not enhance the recall of information more than blocked practice. Still, the latter study did find that interleaving multiple related sets of information did enhance the ability to transfer learning to more integrated problems after a delay. In light of that benefit, it is also important to observe that the study by Richland et al. (14) did investigate scientifically relevant astronomy concepts. Therefore, there is some evidence that the practice of interleaving can provide a benefit to those that are learning physiological definitions and concepts. Nevertheless, the specific nature of that benefit and whether or not interleaving is actually an effective strategy for enhancing the retention of information have yet to be fully elucidated.

Another desirable difficulty that has the potential to benefit physiology education is expanding retrieval practice. This learning strategy is itself linked to two very robust learning theories known as the “testing effect” and the “spacing effect.” The testing effect acknowledges that tests not only assess learning but actually strengthen learning by prompting the recall (i.e., the active processing) of information (2, 5, 9, 10, 17). Furthermore, testing, even without feedback, can increase long-term retention more effectively than repeated studying. The spacing effect indicates that, to a point, the greater the length of time between initial learning and later recall of a piece of information (e.g., testing), the stronger it will be encoded in memory and the more likely it will be retained for a long period of time (1, 2, 4, 5, 6, 9, 10, 12, 13, 16, 17). However, if too much time elapses between initial learning and a recall test, then the information will not be successfully retrieved and will likely be lost from memory. On the other hand, if the time between initial learning and a recall test is too short, the retrieval will be successful but the strategy will fail to effectively promote long-term retention of the information. Thus, if tests are delayed either too little or too long, the strategy will not optimize learning and long-term retention. One way of dealing with this complexity is to use expanding schedules of retrieval tests. With this technique, a test is given immediately after initial learning, which better ensures suc-
cessful retrieval, and then subsequent recall tests occur during systematically increased intervals, which deepen the level of cognitive processing and lead to better long-term retention.

Although the initial support for expanding retrieval practice dates back to 1939 (19), the strategy is most associated with the work of Dr. Bjork and, in particular, with a study he published with Landauer in 1978 (12). In that study, delayed recall of fictitious names was significantly better after an expanding schedule of testing compared with a uniform schedule of testing. Much of the subsequent support for expanding retrieval practice came from studies of clinical populations, including those with a variety of cognitive impairments (16) and those with Alzheimer’s disease (4). However, both of those latter two studies investigated only the effects of expanding retrieval practice, and so neither presented evidence that the strategy is more effective than uniform retrieval practice. In contrast, at least six recent studies that did compare the two different retrieval practices either failed to find a benefit of expanding retrieval (2, 5, 10, 13) or found that uniform retrieval resulted in superior learning and long-term recall (6, 9). On the basis of those types of findings, a recent review (1) of expanding retrieval practice observed that there is actually little evidence supporting an advantage of the strategy and that long-term retention may be more closely related to the average amount of spacing, rather than the specific intervals, between tests.

In light of the controversy described above, Dr. Bjork, along with Storm and Storm (17), published a study in 2010 that not only provided strong support for an advantage of expanding retrieval practice but also provided an explanation as to why many of the above pertinent studies did not find that benefit. According to their study, expanding retrieval practice is only beneficial when the information being learned is vulnerable to forgetting. It is important to point out that, in their study, as with many of the previous similar investigations (2, 5, 9, 12, 13, 17), expanding and uniform retrieval practices were compared using recall tests that were all conducted over a fairly short period of time (e.g., 30–60 min). However, Storm et al. (17) were able to induce forgetting by including an effective distraction task between each of the retrieval tests. Accordingly, they found that expanding retrieval practice resulted in significantly better recall of educational passages on every retrieval test, including a posttest that was administered 1 wk after the information was initially learned. The authors concluded that the previous studies that failed to find an advantage of expanding retrieval practice may have done so because their protocols did not induce sufficient forgetting of the to-be-learned information. They also concluded that another way to induce sufficient forgetting could be to spread the retrieval tests out over longer periods of time (e.g., days or weeks).

In summary, to the best of this author’s knowledge, no study that has spread the retrieval tests out over the course of days has actually shown an advantage of expanding retrieval practice. Furthermore, as was described above, it has also yet to be determined if interleaving is an effective strategy for enhancing the recall of information. Therefore, the purpose of this investigation was to examine the effects of interleaving and expanding retrieval practice on the recall of physiology concepts over a roughly 2-wk period.

METHODS

Students and course. All experimental procedures were approved by the Institutional Review Board of the University of Florida. The students that participated in this investigation were from a fall 2010 APK 2105 Applied Human Physiology course at the University of Florida. The typical Applied Human Physiology student was either a second- or third-year student and was enrolled in a prehealthcare major (e.g., premedicine, prenursing, prepharmacy, prephysical therapy, exercise physiology, etc.).

Experimental procedures. Participants first completed a descriptive questionnaire in which they reported both their biology and chemistry/biochemistry course experience and their specific level of experience with the two experimental topics: immunology and reproductive physiology. Participants were matched according to 1) whether or not they had completed any introductory biology courses, 2) whether or not they had completed any introductory chemistry/biochemistry courses, 3) their self-reported knowledge of the reproductive system (e.g., extensive, some, or none), and 4) their self-reported knowledge of the immune system (i.e., extensive, some, or no knowledge). Participants from each cumulative matched group were then randomly assigned to one of the following four experimental groups: 1) blocked-uniform retrieval (B-UR), 2) interleaved-uniform retrieval (I-UR), 3) blocked-expanding retrieval (B-ER), and 4) interleaved-expanding retrieval (I-ER).

Every participant completed all six sequential phases of the experiment over a 12-day period. The first two phases required participants to read and then reread sets of sentences that summarized an overview of both the immune and reproductive systems. During the latter four phases, participants were repeatedly assessed, without feedback, to prompt them to repeatedly recall the immune and reproductive physiology information they had read. Participants performed all six of the experimental phases using the University of Florida’s E-Learning course management system in Sakai. Briefly, with all six phases, participants first logged on to the E-Learning course website and then selected a link that accessed the appropriate assignment. The individual elements that comprised each of the assignments were then automatically administered by the website, including the instructions for each phase, the reading passages in the first two phases, and all of the assessments in the latter four phases. The E-Learning system was also programmed to help ensure that the participants in each of the four experimental groups completed their assignments on the exact specified days (as described below).

The goal of the first two phases of the experiment was for participants to study the experimental concepts in roughly the same manner in which they normally internalized information in APK 2105. That is, students were generally expected to prepare for lectures by reading ahead in the textbook, and they were then exposed to much of that same information again during a subsequent class meeting and PowerPoint lecture. With that in mind, all the participants began the experiment by reading a total of 17 short paragraphs that summarized numerous key immune and reproductive physiology concepts. Examples of these paragraphs are shown in Table 1. Participants were instructed to carefully concentrate on, and try to retain, the things they read. Those in the two blocked experimental groups read the paragraphs in a blocked manner, meaning that they first read all eight paragraphs pertaining to the immune system in sequential order and then they read all nine paragraphs pertaining to the reproductive system in sequential order. Conversely, those in the two interleaved groups read the paragraphs in an interleaved order (e.g., immune1, reproductive1, immune2, reproductive2, immune3, reproductive3, immune4, etc.). The very next day, all participants completed the second phase of the experiment, in which they reread many of the same immune and reproductive physiology concepts. The author was concerned that the volume and complexity of the experimental passages would make it very difficult for the participants to retain the information throughout the experiment, so the purpose of this second phase
Table 1. Two of the seventeen paragraphs used in the first phase of the experiment

Introduction to the Immune System

The immune system includes leukocytes, lymphoid tissues, and the molecules they produce. The functions of the immune system include clearing foreign material from the body and bringing about long-term immunity to infectious agents. There are five major types of leukocytes: basophils, eosinophils, macrophages, monocytes, and neutrophils.

The Female Reproductive System: Part I

The female reproductive system includes external genitalia, ovaries, and the reproductive tract (uterus, uterine tubes, and vagina). The menstrual cycle lasts about 28 days, includes cyclic changes in ovarian and pituitary hormone secretions, and includes menstruation. Ova develop from germ cells whose number is fixed at birth, and they fully mature only after being fertilized.

Table 2. Four of the thirty sentences used in the second phase of the experiment

- The immune system includes leukocytes, lymphoid tissues, and the molecules they produce.
- There are five major types of leukocytes: basophils, eosinophils, macrophages, monocytes, and neutrophils.
- The menstrual cycle lasts about 28 days, includes cyclic changes in ovarian and pituitary hormone secretions, and includes menstruation.
- Ova develop from germ cells whose number is fixed at birth, and they fully mature only after being fertilized.
37.45 ± 16.78, 49.42 ± 20.57, and 45.11 ± 22.25 for the B-UR, I-UR, B-ER, and I-ER groups, respectively. There was no main effect with the blocked and interleaved conditions \((F = 0.24 \text{ and } P = 0.63)\), nor was there a significant interaction between the four groups \((F = 1.06 \text{ and } P = 0.30)\). There was a significant main effect with the uniform and expanding retrieval conditions \((F = 13.88 \text{ and } P = 0.00)\), the partial \(\eta^2\) was 0.07, and mean scores for those two conditions were 36.75 ± 16.99 and 47.37 ± 21.39, respectively.

All four groups completed the third recall assessment on the fifth day of the experiment. Mean assessment 3 scores were 37.32 ± 18.09, 37.55 ± 18.31, 49.04 ± 22.56, and 45.64 ± 22.43 for the B-UR, I-UR, B-ER, and I-ER groups, respectively. There was no main effect with the blocked and interleaved conditions \((F = 0.28 \text{ and } P = 0.60)\), nor was there a significant interaction between the four groups \((F = 0.37 \text{ and } P = 0.55)\). There was a significant main effect with the uniform and expanding retrieval conditions \((F = 10.87 \text{ and } P = 0.00)\), the partial \(\eta^2\) was 0.06, and mean scores for those two conditions were 37.44 ± 18.11 and 47.42 ± 22.45, respectively.

All four groups completed the final assessment on the 12th day of the experiment, which was 10 days after they had read the concepts in the PowerPoint slideshow. Mean scores for that assessment were 39.76 ± 17.14, 41.12 ± 17.35, 47.98 ± 19.51, and 47.13 ± 20.34 for the B-UR, I-UR, B-ER, and I-ER groups, respectively. There was no main effect with the blocked and interleaved conditions \((F = 0.01 \text{ and } P = 0.93)\), nor was there a significant interaction between the four groups \((F = 0.17 \text{ and } P = 0.69)\). There was a significant main effect with the uniform and expanding retrieval conditions \((F = 6.79 \text{ and } P = 0.01)\), the partial \(\eta^2\) was 0.04, and mean scores for those two conditions were 40.50 ± 17.17 and 47.57 ± 19.81, respectively.

**Table 3. Selected questions from the recall assessment**

1. Leukocytes develop to maturity in the central lymphoid tissues, which include bone marrow and the ________.
2. B cells and T cells provide for all of the following features of immune responses except:
   - A. Specificity
   - B. Diversity
   - C. Memory
   - D. Tolerance
   - E. All of the above are true

3. Which of the following is/are true concerning the menstrual cycle?
   - A. It typically lasts about 28 days.
   - B. It is marked by cyclic changes in ovarian hormones.
   - C. It is marked by cyclic changes in pituitary hormones.
   - D. All of the above are true.
   - E. Answers A and B are true.
   - F. Answers A and C are true.

4. Most cells of the body contain _____ set(s) of chromosomes, but gametes (sperm cells and ova) contain _____ set(s) of chromosomes.
In terms of repeated-measures effects, there was a significant difference in scores across the four assessments ($F = 4.79, P = 0.00$, and $\eta^2 = 0.03$). There was also a significant within-subject interaction between the uniform and expanding retrieval conditions ($F = 10.49, P = 0.00$, and $\eta^2 = 0.05$), but there was not a significant within-subject interaction between the blocked and interleaved conditions ($F = 1.13, P = 0.33$).

**DISCUSSION**

*Summary of main findings.* The purpose of this study was to investigate the effects of interleaving and expanding retrieval practice on the recall of physiology concepts over a roughly 2-wk period. The major conclusions were that 1) there was no advantage to interleaving versus blocked acquisition when learning the physiology concepts but 2) expanding retrieval practice did result in significantly better recall of those concepts up through at least 10 days after initial learning and 1 wk after the final recall assessment.

*Interleaving versus blocked acquisition.* Interleaving has been shown to benefit motor (18), baseball batting (8), and badminton (7) skills as well as the ability to both perform math problems (15) and recognize specific artistic styles (11). However, none of those skills involve the exact types of processing and cognition that are associated with learning physiology information. To the best of the author’s knowledge, only one study (14) has investigated the effects of interleaving on learning scientific concepts, but it is also worth considering that another study (3) examined the strategy with word associations. Although both of those latter studies also found a benefit of interleaving compared with blocked acquisition, the results of the two investigations were somewhat incompatible. More specifically, the study by Battig (3) provided evidence that interleaving increased the recall of learned information, but the more germane study by Richland et al. (14) found that the technique did not increase recall. The results of this investigation therefore agree with those of Richland et al. (14) showing that interleaving acquisition does not increase recall of scientific concepts. Another similarity between this study and that by Richland et al. was that the assessments contained questions that evaluated not only retention (e.g., fill in the blank questions) but also the transfer of learning to more integrated concepts (using complex multiple-choice questions). Whereas Richland et al. found clear evidence of a benefit of interleaving on the transfer of learning, this study did not demonstrate any such benefit. This important discrepancy could be explained by two potential limitations to this study. First, the author does not have a strong background in educational theory and may not have developed assessment questions that properly evaluated the transfer of learning. Second, the participants in this study were asked to learn a fairly large volume of complex words and concepts, and so any potential benefits of interleaving may have been counterbalanced by the difficulty of the task. That is, it is possible that the participants were overwhelmed with information and that interleaving would have been more advantageous if they had been asked to learn less information. However, the author did anticipate the possibility of this problem and consequently structured the methods such that participants were given two opportunities to acquire the experimental concepts (i.e., first in paragraph form and second in individual PowerPoint slides). The above being said, this investigation is one of a very few, if not the only study, that found no benefit of interleaving over blocked acquisition. Future studies in this area should compare interleaved versus blocked acquisition using a smaller volume of physiology information.

*Expanding versus uniform retrieval practice.* In terms of the comparison between retrieval practices, expanding retrieval practice resulted in superior performance on every recall assessment, including the final assessment that was administered after a 1-wk delay. It is important to note that the difference in mean assessment scores between the two conditions significantly decreased with each subsequent assessment, dropping from 16.58 on assessment 1 to 7.07 on assessment 4. It has been suggested that the advantage of expanding retrieval may begin to decrease once retrieval practice ceases (17), and it was certainly not surprising that the expanding retrieval condition assessment scores declined between assessments 1 and 4 (i.e., their assessment 2–4 means were all 6 points lower than their assessment 1 mean). What was very surprising was that the uniform retrieval condition scores actually increased from an average of ~37 on assessments 1–3 to nearly 41 on assessment 4. Given that assessment 4 was administered 1 wk after recall assessment 3 and 10 days after (re)learning the experimental concepts, it is not at all clear why the uniform retrieval group scored significantly higher on assessment 4 than they had on assessments 1–3. Nevertheless, it is very likely that the unexpected increase in their scores, which accounted for one-third of the reduction in the difference across the four assessments between the expanding and uniform conditions, occurred because of some form of experimental error and was not meaningful.

The above being said, the results of this study clearly indicate a benefit of expanding retrieval practice, and so this investigation supports the conclusions of Landauer and Bjork (12) and Storm et al. (17) showing that expanding retrieval practice promotes better learning and retention than does uniform retrieval practice. This study also disagrees with the other similar previous investigations that all found no benefit of expanding retrieval practice after either a short (2, 5, 13) or long (6, 9, 10) delay. As observed by Storm et al. (17), the extent of the benefit of expanding retrieval practice, and consequently the explanation for the discrepancy in the studies mentioned above, may be determined by how vulnerable the to-be-learned information is/was to forgetting. That is, the studies that found no advantage with expanding retrieval practice (2, 5, 9, 10, 13), with the notable exception of that by Cull (6), conducted all of their recall assessments over a short period of time (up to 30–60 min) and likely failed to induce enough forgetting to be able to observe a benefit from the strategy. In contrast, Storm et al. (17) were able to induce sufficient forgetting to observe a benefit by including an effective distraction task between the recall assessments, whereas this study was able to do so by spreading the retrieval assessments out over multiple days. Nevertheless, it is important to point out that Cull (6) also compared retrieval practices using recall assessments that were spread out over several days, which thus should have induced sufficient forgetting, and yet failed to find benefit with expanding retrieval practice. More specifically, Cull examined cued recall of paired associates after either a 1–2–3 day schedule of retrieval assessments.
(expanding condition) or a 2–2–2 day schedule (uniform condition) and found no advantage, and in some cases a disadvantage, with expanding retrieval practice after a 3-day and an 8-day delay.

Other than the type of information that was studied (verbal items vs. physiological definitions and concepts), there were two major distinctions between the methods used by Cull (6) and those used in this investigation. The first distinction was the average spacing between the retrieval assessments, which was only 1 day in this study but 2 days in the study by Cull. A recent review (1) of most of the articles that have investigated retrieval practice suggested that the average spacing between retrieval events may indeed be more important than the actual pattern of spacing. Still, the results reported herein show a clear advantage of expanding retrieval practice, despite having the same average spacing as the uniform condition. It is possible, however, that the extent of the benefit of expanding retrieval practice varies with the actual magnitude of the average spacing between assessments and that Cull observed no such advantage because the average spacing used in that study was twice as great as in this study. Perhaps the benefit of expanding retrieval practice declines as the average spacing between recall assessments increases and/or it disappears when the average exceeds a certain length (e.g., >1 day).

A second distinction that may explain the discrepancy in results between this study and that by Cull (6) was the length of the interval between initial learning and the first retrieval assessment. In this study, those particular intervals were 0 (i.e., immediately) and 1 day in the expanding and uniform conditions, respectively, whereas they were 1 and 2 days in the expanding and uniform conditions, respectively, in the study by Cull. According to Karpicke and Roediger (9, 10), the interval between initial learning and the first recall assessment is more important than the spacing between retrieval events and increases in the length of that interval may significantly improve long-term retention. Similar to the above, although the results of this investigation disagree with Karpicke and Roediger’s conclusions, it is possible that the extent of the benefit of expanding retrieval practice varies with the length of that interval and therefore explains the discrepancies between this study and that by Cull. That is, while expanding retrieval practice produced superior results in this study, despite having a relatively short interval before the first assessment, it is possible that Cull found no such advantage because the operative intervals used in that study were twice as long. In contrast to Karpicke and Roediger’s conclusions (9, 10), perhaps the benefit of expanding retrieval practice declines as the interval preceding the first recall assessment increases and, as Cull (6) actually suggested, the practice is most effective when that interval is no longer than minutes to hours.

Conclusions. In summary, the comparison between retrieval practices reported herein was motivated by the conclusions of Storm et al. (17) showing that expanding retrieval practice can result in superior learning and that the strategy may still be advantageous when the recall assessments are spread out over longer periods of time (e.g., days). This is the first investigation to actually demonstrate an advantage of expanding retrieval practice when the retrieval assessments were conducted over the course of multiple days. Future research in this area should compare expanding and uniform retrieval practices over even longer periods of time (e.g., weeks) and such that the average spacing between assessments is >1 or 2 days. Future studies should also continue to explore the importance of the interval between acquisition and initial recall tests by standardizing the length of that interval between the expanding and uniform retrieval conditions and by varying its duration (e.g., 0, 1, and 2 days).

DISCLOSURES
No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES