The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts

Tracy Linderholm,1 John Dobson,2 and Mary Beth Yarbrough2

1Department of Curriculum, Foundations, and Reading, Georgia Southern University, Statesboro, Georgia; and 2Department of Health and Kinesiology, Georgia Southern University, Statesboro, Georgia

Submitted 26 October 2015; accepted in final form 26 April 2016

The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts. Adv Physiol Educ 40: 329–334, 2016; doi:10.1152/advan.00157.2015.—A testing-based learning strategy is one that relies on the act of recalling (i.e., testing) information after exposure, and interleaving is a strategy in which the learning materials are presented in a serial order (e.g., texts 1, 2, 3, 1, 2, 3, 1, 2, 3) versus a blocked order (e.g., texts 1, 1, 1, 2, 2, 2, 3, 3). Although both learning strategies have been thoroughly investigated, few studies have examined their additive effect with higher-order cognitive tasks such as the ability to identify themes across multiple texts, and none of those did so using physiology information. The purpose of the present study was to compare recall and thematic processing across five different physiology texts. Participants were randomly assigned to learn the texts using one of the following four learning strategies: 1) study-study-study (S-S-S) using a blocked order, 2) S-S-S using an interleaved order, 3) study-test-study (S-T-S) using a blocked order, and 4) S-T-S using an interleaved order. Over the course of the following week, the S-T-S groups had more stable recall of key text ideas compared with the S-S-S groups, and the S-T-S group had more stable recall of thematic information than the S-S-S group when interleaving was used as the presentation order.

Although the majority of empirical work that has been done on the testing effect has addressed fairly superficial levels of learning such as the recall of words or simple concepts, a number of recent studies have found testing advantages with more complex materials and learning tasks (7, 15, 16, 20, 33). The focus of the present study was on the effect of testing on a higher-order cognitive task that has not been investigated, despite the fact that it is often encountered by college-level learners: the ability to read and synthesize ideas across multiple complex science texts.

In college-level courses, students are expected to read and synthesize textbook materials, original research articles, literary novels, etc., to derive common themes. A typical assignment might be to synthesize several original research articles, particularly to set the stage for empirical work that a student may be proposing, to develop a summary of the general findings across multiple studies. Comprehending multiple texts is a complex, multistage task that involves keeping track of the source of ideas, critically evaluating a text and its ideas, integrating and synthesizing ideas, and then making decisions about the importance of ideas within and across texts (24, 25, 32). The ultimate goal of reading multiple texts is to synthesize ideas across texts and store the synthesis in what is called a “documents model” mental representation (32). To provide additional context for what is involved in creating a documents model, the learning process would be akin to the highest levels of learning described by Bloom’s classic taxonomy (6). For example, creating a documents model would involve higher levels of learning such as synthesizing ideas in each text passage, evaluating the importance of the ideas in each text passage, and then creating a new mental model that combines the key ideas from each text passage to identify new themes.

Unfortunately, synthesizing ideas across multiple sources and storing them in an integrated documents model is a daunting task even for college-level students who are advanced in their thinking and reading skills (8, 32). To highlight possible causes for this difficulty, one line of research has indicated that college-level students may incorrectly perceive the purpose of the task of reading multiple text sources incorrectly or have a naïve view of what it means to learn. Specifically, instead of seeing it as an opportunity to transform the text ideas in to new knowledge, they perceive the task as simply relaying the authors’ messages nearly verbatim (8). In addition to perhaps naive beliefs about the nature of learning, it is simply a cognitively complex task that involves many stages (24), which may cause inaccurate processing and comprehension. The end result is that learners tend to build mental models, or memories, of text ideas in isolation and fail to see how ideas across texts connect with one another. That is,
readers fail to build a document model when reading multiple text sources.

Why does the testing strategy facilitate long-term retention and how might it be a strategy that allows for learning a higher-order task such as identifying themes across multiple physiology texts? One framework that has been used to understand how a collection of strategies may facilitate long-term learning attributes their success to the creation of “desirable difficulties” (3, 4). Some learning strategies may actually be more difficult and require more cognitive effort, and this effort creates desirable difficulties. The extra effort required to implement the strategy may lead to elaborative processing, which creates stronger pathways to where information is stored in long-term memory. The benefit is that the stronger memory traces then allows information to be easily searched and retrieved and less relevant information to be suppressed (35). Creating desirable difficulties, or requiring a task that places more demands on limited cognitive resources, at the learning acquisition phase has been shown in numerous implementations to enhance long-term learning of materials (see Ref. 5 for a comprehensive discussion).

Because multiple text comprehension involves necessarily thematic processing and, thus, involves more than long-term recall of simple concepts, an additional strategy that is thought to create desirable difficulties was added to the implementation of a testing strategy in the present study to determine if it would be beneficial. We selected an interleaving text presentation format due to research that shows that it has benefits for higher-order thinking (2, 27, 31, 34). Interleaving refers to presenting materials in a serial order (e.g., texts 1, 2, 3, 1, 2, 3, 1, 2, 3) versus presenting materials in a blocked order (e.g., texts 1, 1, 1, 2, 2, 2, 3, 3, 3). Interleaving is thought to demand additional cognitive effort because it causes contextual interference by creating temporal distance between presentations and can create interference between to-be-learn materials (5, 9). The interference may then promote a compare/contrast mechanism (34), which is essential for identifying themes across multiple text sources. Thus, a secondary objective of the present study was to learn if there are additive benefits to using both a testing strategy and interleaving presentation format for enhancing a higher-order cognitive task such as identifying themes across multiple physiology texts.

One additional challenge that is relevant to using strategies that are perhaps counterintuitive or appear to take more cognitive effort is the role that metacognitive awareness plays in the process. For example, college-level students often have misconceptions about the benefits of strategies that they use for long-term, stable learning (13, 19). For example, most college-level students engage in “cramming,” otherwise known as massed practice, before an exam with the thought that the recency of their memory will help them perform well. Although massed practice may sometimes show immediate benefits, it is not a good strategy for deeper, long-term understanding of course content (19). To understand how metacognition, or one’s assessment of learning, plays a role in the perceptions of the variety of strategies used in this experiment, we also gathered learners’ assessments of and confidence in their learning. Understanding how learners’ awareness of the benefits of using strategies that may create desirable difficulties would add a new dimension to this body of work and help college instructors realize the challenges that may go along with their recommendations to use strategies that create desirable difficulties.

To explore the questions outlined above, there were three major objectives identified for the present study. First, we sought to determine the benefits of testing for a more complex cognitive task such as comprehending themes across multiple physiology text passages from different sections of a popular undergraduate textbook. Second, because thematic processing across multiple texts is so challenging even for advanced students, we investigated whether or not adding an interleaving presentation format would have an additive benefit to students’ learning. The final objective was to determine if metacognitive awareness, as measured by confidence in one’s learning, is related to the use of different strategies for reading multiple physiology texts.

METHODS

Participants

Students enrolled in a physiology course at a regional university in the southeastern part of the United States were offered a small amount of course credit for participating in the study. The typical student in this course was a fourth-year student and was either an Exercise Science or Athletic Training major. All participants were native English speakers. The university’s Institutional Review Board approved all experimental procedures.

Design

The design used in this experiment was between subjects, and there were four different groups of learning conditions. Participants were randomly assigned to learn multiple texts using either a study-study (S-S-S) strategy or a study-test-study (S-T-S) strategy that had either an interleaved or blocked text presentation format. The participants’ ability to process multiple texts thematically and to identify ideas from individual text passages was determined both immediately after learning and then again after a 1-wk delay. The participants’ confidence levels in their learning both prior to and after creating their immediate set of essay responses were gathered to serve as a reflection of their metacognitive awareness.

Materials

Texts. To understand how participants were able to synthesize concepts and identify themes across multiple texts, all participants read five expository text passages that were modified from an undergraduate-level college textbook, Exercise Physiology, by Powers and Howley (Ref. 26, p. 265, 267, 313, 314 and 356). The text passages had a common concept of maximal oxygen consumption ($V_{O_{2 max}}$) and the various factors that influence it or its measurement. The titles of the five text passages were as follows: “Heritage study,” “Why do some untrained individuals have high $V_{O_{2 max}}$ values,” “$V_{O_{2 max}}$ and the plateau,” “Errors in estimating $V_{O_{2 max}}$,” and “Changes in $V_{O_{2 max}}$ with age in women.” The average length of the five text passages was 12.8 sentences (range: 11–14 sentences/text) and 319.20 words (range: 291–362 words/text).

Confidence assessments. Each participant was asked to rate their confidence on their understanding of each text and their ability to remember information from each text on a Likert-type scale (0%, 20%, 40%, 60%, 80%, and 100%), where 0% indicated “not confident at all” and 100% indicated “extremely confident.” At the first session only, participants were asked to complete the confidence assessment right after reading the texts (time 1) and then again after completing the essay task (time 2). Three following questions were asked with regard to each of the five texts at time 1: question 1 was “How certain are you that you understood the main points of the text: _____?"
question 2 was “How certain are you that you will recall the text: ______ in a week?” and question 3 was “How certain are you that you will recall the text: ______ in 5 min?” At time 2, only questions 2 and 3 were asked of participants. Thus, at time 1 participants completed 15 confidence questions and at time 2 participants completed 10 confidence questions.

Essay task. Participants were asked to create a short essay about their conceptual understanding of the five expository texts and were encouraged to write broadly about themes. The directions to participants were presented in writing for consistency and were as follows: “Please compare/contrast the five texts’ thematic similarities and differences in writing. Be concise, organize your ideas in a logical sequence, and focus on themes or key ideas that were present in and across texts rather than writing down all the isolated facts you encountered. Please write legibly and clearly. There will be a 10-min time limit for this task. When the 10 min are up, the researcher will ask you to hand in this task.”

Procedures

In this experiment, there were two sessions that occurred in group settings. Session II occurred 1 wk after session I.

Session I. Participants signed in with the research assistant and were randomly assigned to one of the four learning conditions. Participants completed a consent form and were given a packet of the expository text materials that aligned with their condition. The instructions on the packets were as follows: “Each page in this packet will have one task for you to complete. Please follow the instructions at the top of each page to complete the task. You will have 2 min to complete each task. The researcher will indicate to you when it is time to move to the next task. Please do not begin the first task until the researcher tells you to do so.” At the top of the pages where participants were to read or reread a passage, the instructions read as follows: “Please read the following passage for comprehension.” On the top of the pages where participants were to engage in self-testing, the directions read as follows: “Please write in the space below everything you can remember about the text titled, ______.”

There were three versions of each condition in the experiment that varied the presentation order of the texts to ensure that order effects did not occur. In the S-S-S/interleaved condition, participants had three study phases where they read each of five texts and then repeated the sequence two times. In the S-T-S/interleaved condition, participants read each of five texts in a row, recalled each of the five texts, and then reread each of the five texts. In the S-S-S/blocked condition, participants read the first text first and then reread it twice before moving on to the next of the five texts. Finally, in the S-T-S/blocked condition, participants read the first text in the series, recalled it, and then reread it before moving on to the next text. They repeated this pattern for each of the five texts. As is typical of studies that have compared testing with other strategies (11, 12, 14, 29, 30), the total duration of learning was standardized by giving the participants 2 min to complete each task within each of the four learning strategies. Before data collection, two graduate students studied all five of the texts and helped the investigators determine that the 2-min interval was sufficient to perform each learning task

Upon completion of the study phase of the experiment, packets were collected by the research assistant and the participants then completed the first confidence assessment. After confidence assessments were collected, participants then completed the essay task. Participants then completed the second confidence assessment. Once completed, participants were reminded to come back to their second experimental session and were not to discuss the session with classmates. Session I took participants on average 1 h.

Session II. Participants came back 1 wk later and completed the essay task a second time. After handing in the task, participants were asked to complete a demographic questionnaire. Finally, they were debriefed about the nature of the study in writing and were excused. Session II took participants an average of 20 min to complete.

Scoring

The essay task was scored by two of the coauthors. Before scoring, the three coauthors, two of whom are content experts, determined three overarching themes that cut across all five texts. The authors then identified key ideas from the texts, and this was determined to be 16 in total. To weight the scores to emphasize thematic processing and synthesis, five points were given to participants for each key theme that was present in essays. One point was given for each of the 16 key ideas identified in the essays. For each theme and key idea, participants were assigned either zero points, half-credit points, or full credit points depending on the completeness and accuracy of information contained in the essays. Thus, for key themes a total of 15 points were possible and for key ideas a total of 16 points were possible.

Once the rubric was developed, two of the coauthors trained on five essays to ensure the rubric was appropriate and that each of us understood the scoring task. The five essays were then returned to the set for final scoring. To determine agreement levels, the two coauthors then selected 20 essays (10 participants who wrote 2 essays each) to score independently. None of the five essays that were used for training purposes were used to determine agreement level. The individuals who scored participants’ data were blinded to the condition to which participants were randomly assigned. An agreement level of 93.07% was reached for the 20 essays scored. Disagreements were resolved through discussion. Given that the agreement level was deemed sufficient, one coauthor then scored the remaining essays independently.

Data Analysis

Data pertaining to the ability to recall and synthesize ideas from multiple physiology texts were analyzed using repeated-measures ANOVAs, and followup comparisons were conducted using t-tests. Effect sizes were expressed using partial eta squared and Cohen’s d for the ANOVAs and t-tests, respectively. Multivariate analysis of variance and bivariate correlations were used to analyze participants’ confidence levels. SPSS (version 21, IBM, Armonk, NY) was used to perform statistical analyses. Statistical significance was set at \( P < 0.05 \). Scores are expressed as means ± SE.

RESULTS

A total of 95 undergraduate students volunteered to participate in the study. The average age of participants was 21.6 yr, and there were 64 female participants and 31 male participants in the sample. Two participants were removed from the sample for failure to follow instructions on the essay portion of the task. Of the 93 remaining participants, 23 participants were in the interleaved/S-S-S condition, 24 participants were in the interleaved/S-T-S condition, 21 participants were in the blocked/S-S-S condition, and 25 participants were in the blocked/S-T-S condition.

Recall of Cross-Cutting Themes in Written Essays

Given that the primary interest was in how different presentation formats and strategies facilitated participants’ ability to recognize themes across texts, essay responses that indicated an understanding of themes that cut across the five exercise science texts were inspected first. Those responses were evaluated using 2 (presentation format: interleaved or blocked) × 2 (learning strategy: S-S-S or S-T-S) × 2 (time: immediate and 1-wk test) repeated-measures ANOVA. The results indicated that a main effect of time was found \( F(1, 89) = 17.42, P < \)
0.16] as well as a two-way interaction between learning strategy and time \(F(1, 89) = 4.89, P < 0.05, \eta^2_p = 0.05\). However, these effects were qualified by a significant three-way interaction between presentation format, learning strategy, and time \(F(1, 89) = 7.83, P < 0.01, \eta^2_p = 0.08\); Table 1. The two-way interaction was further analyzed next. The breakdown of the three-way interaction was first inspected from the perspective of presentation format. First, the interleaving condition was analyzed in isolation. There was a simple main effect of time \(F(1, 45) = 7.85, P < 0.01, \eta^2_p = 0.15\). But, this was qualified by a simple interaction between learning strategy and time \(F(1, 45) = 11.77, P < 0.01, \eta^2_p = 0.21\). Followup comparisons revealed that the S-S-S condition had a significant decrease from the immediate test \((M = 5.44 \pm 0.44)\) to the delayed test \([M = 3.37 \pm 0.45, t(22) = 3.69, P < 0.01, d = 1.57]\). For the S-T-S condition, thematic recall held steady and was no different at the immediate test point \((M = 4.17 \pm 0.44)\) versus the delayed test point \((M = 4.38 \pm 0.44, t < 1)\). At the immediate test point, there was a significantly higher thematic recall for S-S-S \((M = 5.44 \pm 0.44)\) versus S-T-S \([M = 4.17 \pm 0.44, t(45) = 2.18, P < 0.05, d = 0.65]\). The difference between conditions did not remain at the delayed test point \((t < 1)\). With regard to the blocked condition, there was only a main effect of time \(F(1, 44) = 9.74, P < 0.01, \eta^2_p = 0.18\). No other effects reached significance.

The breakdown of the three-way interaction was then inspected from the perspective of learning strategy. First, when effects were isolated relative to the S-S-S condition, there was only an effect of time \(F(1, 42) = 16.32, P < 0.01, \eta^2_p = 0.28\). No other effects reached significance. Second, when effects were isolated relative to the S-T-S condition, there was an interaction between presentation format and time \(F(1, 47) = 5.26, P < 0.05, \eta^2_p = 0.10\). Followup tests revealed a significantly larger level of thematic recall in the blocked condition at the immediate test point \((M = 4.90 \pm 0.51)\) than the interleaving condition \([M = 4.17 \pm 0.36, t(24) = 2.53, P < 0.05, d = 1.03]\). No other effects reached significance.

The breakdown of the three-way interaction was then inspected from the perspective of time. Focusing first on the immediate test point, analyses revealed no simple main or interactive effects between the presentation format or learning strategy conditions \((F_S < 1)\). Likewise, for the delayed test point, there also were no simple main or interactive effects for these variables on thematic essay responses \((F_S < 1)\).

### Table 1. Recall of themes and key ideas as a function of learning strategy, presentation format, and time

<table>
<thead>
<tr>
<th>Essay themes*</th>
<th>Immediate</th>
<th>1 Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-S-S/interleaved</td>
<td>5.44 ± 0.44</td>
<td>3.37 ± 0.45</td>
</tr>
<tr>
<td>S-T-S/interleaved</td>
<td>4.17 ± 0.44</td>
<td>4.38 ± 0.44</td>
</tr>
<tr>
<td>S-S-S/blocked</td>
<td>5.12 ± 0.47</td>
<td>4.29 ± 0.47</td>
</tr>
<tr>
<td>S-T-S/blocked</td>
<td>4.90 ± 0.43</td>
<td>3.80 ± 0.43</td>
</tr>
<tr>
<td>Key ideas(^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-S-S/interleaved</td>
<td>2.98 ± 0.38</td>
<td>1.59 ± 0.31</td>
</tr>
<tr>
<td>S-T-S/interleaved</td>
<td>2.19 ± 0.37</td>
<td>1.73 ± 0.30</td>
</tr>
<tr>
<td>S-S-S/blocked</td>
<td>2.62 ± 0.39</td>
<td>1.69 ± 0.32</td>
</tr>
<tr>
<td>S-T-S/blocked</td>
<td>2.50 ± 0.36</td>
<td>2.14 ± 0.29</td>
</tr>
</tbody>
</table>

**Recall of Key Ideas in Written Essays**

A secondary interest in the study was to determine if presentation format or learning strategy conditions would have singular or additive effects on recall of key ideas encountered in the five texts. That analysis was also performed using 2 (presentation format: interleaved or blocked) \(\times\) 2 (learning strategy: S-S-S or S-T-S) \(\times\) 2 (time: immediate and 1-wk test) repeated-measures ANOVA. The results revealed a significant effect of time \(F(1, 89) = 19.09, P < 0.01, \eta^2_p = 0.18\). But, the main effect was qualified by an interaction between learning strategy and time \(F(1, 89) = 4.37, P < 0.05, \eta^2_p = 0.05\). Followup tests revealed that recall of key ideas decreased significantly from the immediate test point \((M = 2.80 \pm 0.29)\) to the delayed test point \((M = 1.64 \pm 0.23)\) but only for the S-S-S condition \([t(43) = 4.09, P < 0.01, d = 1.25]\). Recall of key ideas did not differ from the first test \((M = 2.35 \pm 0.24)\) to the second test \((M = 1.94 \pm 0.20)\) for those in the S-T-S condition \((t < 1)\). No other effects were significant.

**Confidence Levels**

To understand how confidence levels varied as a function of presentation format and learning strategy, 2 \(\times\) 2 multivariate analysis of variance was conducted using presentation format (interleaved or blocked) and learning strategy (S-S-S or S-T-S) as independent variables on each type of confidence measure: assessment of understanding (preessay), confidence in recall in 5 min (preessay), confidence in recall in 1 wk (preessay), confidence in recall in 5 min (postessay), and confidence in recall in 1 wk (postessay).

The results showed a main effect of presentation format on one’s confidence in recall in 5 min when ratings were done before the essay task was completed \(F(1, 92) = 4.05, P < 0.05, \eta^2_p = 0.04\). In detail, participants’ confidence ratings were higher in the interleaved condition \((M = 0.58 \pm 0.03)\) compared with the blocked condition \((M = 0.49 \pm 0.03)\). The results also showed a main effect of learning strategy on the assessment of understanding and confidence in recall in 5 min when ratings were done before the essay task \([F(1, 92) = 5.16, P < 0.05, \eta^2_p = 0.06]\; \(F(1, 92) = 5.25, P < 0.05, \eta^2_p = 0.06\), respectively]. In detail, participants were more confident in their understanding of the texts in the S-S-S condition \((M = 0.68 \pm 0.03)\) than in the S-T-S condition \((M = 0.59 \pm 0.03)\). Likewise, participants were more confident in their ability to recall text information in 5 min in the S-S-S condition \((M = 0.59 \pm 0.03)\) than in the S-T-S condition \((M = 0.49 \pm 0.03)\) before the essay task was completed. No other effects reached significance \((F_S < 1)\).

A series of bivariate correlations were performed to learn if confidence measures were related to actual essay task performance. Interestingly, the majority of correlations between confidence measures and recall task components (recall of themes vs. key ideas) were negative and did not reach significance (see Table 2). However, it should be noted that measures of confidence all correlated positively and significantly with one another. For example, confidence in one’s understanding of text ideas before the essay task was correlated with one’s confidence in their ability to recall in 5 min \((r = 0.82, P < 0.01)\) and after 1 wk \((r = 0.57, P < 0.01)\).
Table 2. Bivariate correlations between actual recall and confidence assessments at two test points

<table>
<thead>
<tr>
<th>Actual recall: essay themes Test point 1</th>
<th>Understand</th>
<th>Immediate Recall</th>
<th>I-wk Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test point 1</td>
<td>-0.16</td>
<td>-0.15</td>
<td>-0.04</td>
</tr>
<tr>
<td>Test point 2</td>
<td></td>
<td>-0.16</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actual recall: key ideas Test point 1</th>
<th>Understand</th>
<th>Immediate Recall</th>
<th>I-wk Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test point 1</td>
<td>-0.04</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>Test point 2</td>
<td></td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

DISCUSSION

The three objectives of the present study were as follows: 1) to determine if the testing effect enhances performance for a complex cognitive task such as comprehending multiple physiology texts; 2) to determine whether or not there is an additive benefit to using two strategies that create desirable difficulties (an interleaving presentation format along with a testing strategy for recall of both themes across multiple texts as well as and key ideas from individual texts); and 3) to determine whether or not learners’ levels of confidence in their ability to understand and recall text ideas, which influences their motivation to use appropriate strategies, aligned with strategy condition and performance.

The results with regard to study objectives 1 and 2 showed that when the presentation format involved interleaving, participants recalled more thematic information immediately in the S-S-S condition compared with the S-T-S condition but that recall declined rapidly in the long term for the S-S-S condition. That is, participants’ recall of themes was more stable over a time delay in the S-T-S condition than in the S-S-S condition when interleaving was used as the presentation format. Specifically, participants in the S-S-S condition recalled 38% less thematic information from the immediate test point to the delayed test point, whereas participants in the S-T-S condition actually recalled slightly more at the delayed test point (5%) compared with the immediate test point. Such an effect was not present when a blocking format was used. Thus, we have both evidence that the testing effect is useful for enhancing performance on cognitively complex tasks such as reading and comprehending multiple expository texts and also have evidence that interleaving used with a testing strategy can have additive, stable long-term memory benefits for recalling text themes in some situations. This is a novel finding because the testing and interleaving strategies have yet to be studied on a reading task where college-level learners must retain and synthesize thematic information from multiple, complex physiology texts.

A second set of analyses was performed on the recall of key ideas within each of the five physiology texts. Results from these analyses showed again that recall was more stable over time in the S-T-S condition relative to the S-S-S condition. In the S-S-S condition, participants’ recall significantly declined from the immediate to the delayed test point by 41%, whereas recall declined by 17% in the S-T-S condition. There were no additive benefits for using the interleaving presentation format. So with regard to the recall of key ideas relevant to individual texts and embedding them in one’s written summary of multiple texts, we have evidence that the testing strategy is beneficial irrespective of presentation format. Our tentative conclusion is that the task of recalling key ideas from multiple texts is not complex enough to require the use of two strategies that invoke desirable difficulties.

Finally, participants’ confidence in and awareness of their own understanding of text information and anticipated recall performance was highest for participants, at least initially, in the interleaved and S-S-S conditions. In detail, participants were most confident right after reading the texts (and before the essay task was completed) with respect to believing that they understood text content and that they could recall it well in 5 min. As previously reported, there was not a relation between confidence levels and actual essay performance. Thus, learners were largely unaware of what strategies would yield the best performance with the exception of having some initial sense that the interleaving strategy was beneficial to short-term recall. This is similar to most studies of metacognition that adult learners are often inaccurate at judging their performance (e.g., Ref. 13).

Taken together, the results have both theoretical and practical applications. Theoretically, these results show that the testing effect is robust in that it can be applied to surface recall of simple concepts (7, 16, 23, 28, 29) as well as boost learners’ ability to synthesize and draw themes from several texts that have overlapping concepts. Of course, multiple text processing is enhanced when an interleaved text presentation is present as well but only when something as complex as thematic processing is required. So, for the desirable difficulties framework (3, 4), there is some evidence for an additive effect of multiple strategies when the task is complex enough. For simpler tasks, such as recall of key ideas, perhaps one desirable difficulties strategy, such as self-testing, is sufficient to yield recall benefits.

For physiology and exercise science college instructors, there are some practical applications for the results. Instructors may suggest that materials be interleaved and to use testing when students are studying independently and also model interleaving and testing in exam review sessions by returning to the assessment of concepts after some time has passed. Given the results that showed that confidence ratings and perceptions of understanding do not match precisely with actual performance, instructors can show students data or even perform simulations in class that demonstrate to students which strategies work best for long-term retention. Other researchers have found that when college-level students are shown data of effective strategies in the context of their courses that they are more likely to understand their importance and use them when studying independently (11, 14). This aligns with related evidence that although students may have misconceptions about the best learning strategies it is possible that they can correct their knowledge with the proper intervention (13). But perhaps the most striking application of this study is that learning strategies that create desirable difficulties can be useful when learning more complex physiology concepts or engaging in thematic learning that extends beyond basic vocabulary recall. College-level learners should be encouraged to engage in these types of strategies particularly as they advance toward the end of their undergraduate program and the expectations rise in terms of the complexity of what they are asked to do in their courses and how long they are required to maintain learned information.
DISCLOSURES

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

AUTHOR CONTRIBUTIONS

T.L. and J.L.D. conception and design of research; T.L. and M.B.Y. performed experiments; T.L., J.L.D., and M.B.Y. analyzed data; T.L. and J.L.D. interpreted results of experiments; T.L. and J.L.D. drafted manuscript; T.L., J.L.D., and M.B.Y. edited and revised manuscript; T.L., J.L.D. and M.B.Y. approved final version of manuscript.

REFERENCES