ASE

The Effect of Selected "Desirable Difficulties" on the Ability to Recall Anatomy Information

John L. Dobson,[1]* Tracy Linderholm[2]
[1]*Department of Health and Kinesiology, Georgia Southern University, Statesboro, Georgia*
[2]*Department of Curriculum, Foundations, and Reading, Georgia Southern University, Statesboro, Georgia*

"Desirable difficulties" is a theory from cognitive science used to promote learning in a variety of contexts. The basic premise is that creating a cognitively challenging environment at the learning acquisition phase, by actively engaging learners in the retrieval of to-be-learned materials, promotes long-term retention. In this study, the degree of desirable difficulties was varied to identify how cognitively challenging the learning acquisition phase must be to benefit university-level students' learning of anatomy concepts. This is important to investigate as applied studies of desirable difficulties are less frequent than laboratory-based studies and the implementation of this principle may need to be tailored to the specific field of study, such as anatomy. As such, a read-read-read-read (R-R-R-R) condition was compared to read-generate-read-generate (R-G-R-G) and read-test-read-test (R-T-R-T) conditions. The three conditions varied in terms of how effortful the retrieval task was during the learning acquisition phase. R-R-R-R required little effort because participants passively read the materials four times. R-G-R-G required some effort to generate a response as participants completed a word fragment task during the learning acquisition phase. R-T-R-T was thought to be most demanding as participants performed a free recall task twice during the learning phase. With regard to the absolute amount of anatomy information recalled, the R-T-R-T condition was superior at both immediate and delayed (one week) assessment points. Thus, instructors and learners of anatomy would benefit from embedding more free recall components, or self-testing, into university-level course work or study practices. Anat Sci Educ 8: 395–403. © 2014 American Association of Anatomists.

Key words: gross anatomy education; allied health education; undergraduate education; desirable difficulties; testing; retrieval practice; generation effect; learning; knowledge recall

INTRODUCTION

The focus of this study is to determine how to most effectively apply laboratory-based learning phenomena, that have been well-documented by cognitive scientists, to boost university-level student learning in the allied health professions. Previous work in this area has focused on applying cognitive science findings to enhance learning in medical education and anatomy and physiology course content (e.g., Larsen et al., 2009; Logan et al., 2011; Larsen et al., 2013; Dobson and Linderholm, 2015). The specific objective of this study is to further extend previous investigations that have capitalized on what cognitive scientists call the "testing effect" to enhance student learning of difficult anatomy materials. Prior investigations have demonstrated that asking university-level students to actively attempt to recall anatomy and physiology materials during the learning acquisition phase yields superior retention compared to simply reviewing or rereading the materials prior to final memory assessments (Dobson and Linderholm, 2015). The unique component of the current study is to determine to what degree students must be engaged in the retrieval process in order to see benefits to long-term learning. Some learning strategies may involve more effort and involvement on the part of the learner. For example, a free recall task, the task most often used in testing effect studies, provides the learner with no

*Correspondence to: Dr. John Dobson, Department of Health and Kinesiology, Georgia Southern University, P.O. Box 8076. Statesboro, Georgia 30460. E-mail: jdobson@georgiasouthern.edu

Received 5 May 2014; Revised 2 July 2014; Accepted 21 August 2014.

Published online 16 September 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/ase.1489

memory cues and so may be considered the most effortful and engaging if the goal is to recall accurately. In the current study, a free recall task is contrasted against a word fragment task where learners are given some cues about the correct answer to recall. The completion of a word fragment task, with memory cues and a built-in feedback mechanism, may require less effort and engagement on the part of the learner. Another unique element of the study is that some researchers propose that the laboratory-based findings of the testing effect may need to be tailored depending on the nature of the to-be-learned information (Richland et al., 2005). This study attempts to refine how to apply this literature to learning groups of anatomy terms that are organized by muscle set. Understanding the nature of the retrieval process that must be used during the learning acquisition phase is important for anatomy instructors to know so that they may identify how much control they have over this learning technique and how to advise students to use it. To meet the objective of the study, three experimental conditions were developed based on variations in the nature of the recall task and based on a continuum of how cognitively demanding the recall task was during the learning acquisition phase.

The experimental conditions in this study were varied based on how extensively they created what cognitive scientist Robert Bjork (Bjork, 1994; Bjork and Bjork, 2011) called "desirable difficulties" which is another way of describing the degree of cognitive demand during the learning acquisition phase. The theory of desirable difficulties can be used to explain why some learning strategies do not necessarily improve learning in the short-term but do so in the long-term (Bjork, 1994; Roediger and Karpicke, 2006). This theory proposes that learning strategies that require additional cognitive resources, and are necessarily more cognitively demanding, may hinder learning in the short run but facilitate stronger long-term memories. These cognitively demanding strategies actually produce more errors during initial learning, causing learners to struggle and perhaps engage in more elaborative processing, but this additional effort typically leads to stronger memories when tested after a delay of up to one week later (e.g., Roediger and Karpicke, 2006; Bjork et al., 2013).

Perhaps the most straightforward examples of two experimental conditions that vary in terms of desirable difficulties are the much investigated and replicated "study-study" versus the "study-test" conditions (e.g., Roediger and Karpicke, 2006; Larsen et al., 2009; Karpicke and Blunt, 2011; Jönsson et al., 2012). In the study-study condition, participants are asked to study a set of to-be-learned materials. In part two of the learning sequence, they are asked to simply study the materials a second time. In the study-test condition, participants are asked to study the materials in the first phase of the sequence. But in part two of the study sequence participants are commonly asked to retrieve, via free recall, all that that they learned from part one of the learning phase. So instead of reviewing or restudying materials, they are asked to perform a retrieval task. Attempting to retrieve from memory all that was learned in part one of the study sequence is certainly more difficult than simply restudying materials. The pattern of results, deemed the testing effect, have consistently shown that in the short-term, the study-test condition is inferior to the study-study condition in terms of proportion of to-be-learned items that are recalled but the study-test leads to superior recall when participants are assessed after a delay of several hours or a week (Roediger and Karpicke, 2006). This pattern of results is in line with the desirable difficulties theory and indicates that the study-test condition creates more errors and effort during the initial learning phase but the effort pays off in the long run with stronger memory traces of the information.

Other theoretical explanations exist to describe the underlying cognitive mechanisms that may be at work to cause the testing effect. Two such examples are the mediator effectiveness hypothesis (Pyc and Rawson, 2010) and elaborative retrieval processing view (Carpenter and DeLosh, 2006). The mediator effectiveness hypothesis (Pyc and Rawson, 2010) purports that testing is a superior learning strategy because it strengthens the encoding of mediators that help learners create links between cues and to-be-retrieved materials. The elaborative retrieval processing view (Carpenter and DeLosh, 2006) claims that the act of retrieval forces elaboration of the to-be-learned materials, making the materials easier to access in memory later on. Successful retrieval appears to be particularly enhanced when there is interference introduced or limited retrieval cues, forcing the learner to elaborate. These alternative theories seem to be compatible with the desirable difficulties framework because all three claim that the benefit of self-testing is that encoding is strengthened by elaborative and effortful processing of to-be-learned materials.

Although the precise underlying cause for the testing effect requires further investigation, the testing effect pattern itself has been replicated consistently and is applicable to both laboratory (e.g., Roediger and Karpicke, 2006) and applied settings (e.g., Glover, 1989; Larsen et al., 2013). The testing effect has been found to apply to a wide variety of age groups (e.g., Lipowski et al., 2013; Meyer and Logan, 2013) and across a variety of disciplines (e.g., Larsen et al., 2008; Logan et al., 2011; Einstein et al., 2012). It appears to be the superior strategy even when compared to other well-established active learning techniques such self-explanation (e.g., Larsen et al., 2013) and is productive in situations where learners are not given feedback about the accuracy of their study strategies (e.g., Butler and Roediger, 2007). Perhaps most encouraging, self-testing can be useful for generalizing learning beyond only the to-be-learned items (Zarombe and Roediger, 2010; Hinze and Wiley, 2011). In other words, engaging in the practice of self-testing, particularly when retrieval of to-be-learned items is complete and open-ended as in free recall tasks (Glover, 1989), can lead to higher order learning and transfer of knowledge to new contexts. Clearly, the testing effect is robust and can be useful to understand or enhance a variety of learning contexts.

Most relevant to the current study, several researchers have applied the testing effect strategy to teaching in the allied health professions (e.g., Logan et al., 2011; Larsen et al., 2013; Dobson and Linderholm, 2015). For instance, Dobson and Linderholm (2015) instructed university-level anatomy and physiology students to study course materials using a study-study-study strategy versus a study-study/take notes strategy versus a study-test-study strategy. The results showed that after a one-week delay, the study-test-study condition was superior to the other two conditions in terms of the proportion of information recalled. After this first phase of the investigation was completed, the students in the section of the anatomy and physiology course who participated in the experiment where shown their collective data and urged to use a testing strategy, similar to the study-test-study strategy condition, when preparing for course exams on their own. Subsequent course exam scores were compared between

sections of the same course. Students who participated in the initial experiment, and who were shown the benefits of testing, had superior course exam performance compared to students in other course sections who did not participate in the experiment. Thus, the testing effect can clearly be used to facilitate long-term learning of anatomy concepts for university-level students.

One essential component of learning anatomical information is that students must be prepared to learn large volumes of terms that are often derived from Greek or Latin for the name and typical composition, appearance, and relative position of the structures of the body. Generation, where learners generate a portion of to-be-learned items, may be an ideal strategy to use for learning anatomical information because it involves retrieval, creates desirable difficulties during the learning acquisition phase, and also appears to be useful for learning facts (see Richland et al., 2005). One example of generation is when participants are given information to study and then they must generate an answer by completing a word fragment task. For example, if participants are asked to learn simple paired associates such as "HOT – COLD," completion of the word fragment "HOT - C_ L _" can be used to determine how quickly and accurately participants can generate the letters from memory. Generation was added as a strategy condition to the current study design given that it might be easier for instructors to manipulate a retrieval strategy by designing word fragment materials and it allows for students to get some kind of feedback if their generated term/terms "fits" the word fragment. In some studies, feedback has been found to be an important component for building long-term memories that are accurate (e.g., Richland et al., 2005). That, together with the finding that generation is useful for simple fact retrieval, motivated the generation condition in the current investigation.

To summarize the design and reiterate the purpose of the current study, our goal was to further explore the extent of how cognitively demanding, or difficult, the retrieval task needs to be in order to have a long-term benefit to the student by varying the nature of the retrieval task in the learning acquisition phase. For this work to be most applicable to use in anatomy courses, it is important to know how extensive desirable difficulties need to be in order to facilitate learning. If this can be precisely determined, it will allow instructors to appropriately advise students how to engage in study strategies and/or it will allow instructors to precisely design learning experiences themselves. Therefore, three learning conditions were created. The conditions varied in terms of how much disruption may take place, during the learning acquisition phase. Condition 1 required participants to carefully read through to-be-learned anatomical materials four consecutive times (R-R-R-R). Condition 1 is akin to the study-study variety of tasks where there is minimal cognitive effort on the part of the participants because the participants simply review the materials. Condition 2 required the participants to read the materials and then generate correct answers from memory by completing a word fragment task. This was repeated two times in a row (R-G-R-G). The word fragment task is a form of a generation task that is perhaps less difficult than unconstrained free recall from memory. Condition 3 required participants to read the materials, free recall, re-read the materials, and then free recall again (R-T-R-T). This condition is akin to the traditional study-test condition and is expected to be difficult for participants given there are no constraints or cues regarding what must be recalled.

METHODS

All experimental procedures were approved by the University's Institutional Review Board. The participants were recruited from a Structural Kinesiology course at a regional U.S. university. The typical Structural Kinesiology student was a third year student and was an allied health or similar major (e.g., pre-physical therapy, exercise science, athletic training, pre-medicine, etc.).

Treatment

Students repeatedly studied the origins, insertions, actions and innervations of three sets of three skeletal muscles. The first set of muscles included the adductor hallucis, lumbricals of the foot, and plantar interossei muscles; the second set included the platysma, occipitofrontalis, and medial pterygoid muscles; and the third set included the scalenes, masseter, and temporalis muscles. These muscles were chosen specifically because they were neither discussed, nor did they appear on any other assignment, throughout the Structural Kinesiology course. All students used each of the following strategies to study the muscle sets, but they were randomly assigned to use only one strategy to study each set. One strategy required students to carefully read through the information in a muscle set four consecutive times (R-R-R-R). Another strategy required them to carefully read a muscle set and then generate the same information from word fragments two consecutive times (R-G-R-G). The word fragments were created by randomly removing one or two letters from each to-be-learned item. Participants attempted to complete the word fragment task twice. One of the purposes of the second attempt at generation was to provide students feedback as to how effectively they had generated that information by letting them see the word fragments again. A third strategy required students to carefully read a muscle set and then immediately test themselves on (i.e., freely recall) that information two consecutive times (R-T-R-T). During the testing portions of the R-T-R-T strategy, students could not see any of the muscle set information, and they were instructed to write down as much as they could recall on a provided sheet of paper. Table 1 displays the read, test and generate conditions that students used when studying one of the muscle sets.

The experimental muscle sets and studying strategies were administered both in sequential order and during only one overall studying session. Each of the four studying conditions within each of the three strategies lasted exactly three minutes. At the conclusion of each three minute condition, students were instructed to move to the next condition or strategy, and they were not allowed to go back and repeat any of the previous conditions or strategies. Therefore, every student spent exactly twelve minutes both studying each muscle set and using each studying strategy. Finally, in an effort to minimize any potential advantages or disadvantages related to using a particular studying strategy at the beginning versus the end of the sequential studying session, the order in which the participants used the strategies was randomized.

Experimental Procedure

All Structural Kinesiology students were required to complete four exams throughout the semester. Following the

Table 1.

The Read, Generate, and Test Study Conditions of One of the Experimental Muscle Sets

Read Condition – Carefully and continuously read through the following muscle information:		
1. Platysma	*Origin:*	Fascia covering superior portion of pectoralis major
	Insertion:	Base of mandible, skin, and lower part of face
	Actions:	Depress mandible. Tighten neck fascia. Draw down corners of mouth
	Innervation:	Facial
2. Occipitofrontalis	*Origin:*	Galea aponeurotica
	Insertion:	Skin superior to eyebrows and superior nuchal line
	Actions:	Raise eyebrows and wrinkle forehead. Retract the galea aponeurotica
	Innervation:	Facial
3. Medial Pterygoid	*Origin:*	Medial surface of lateral pterygoid plate of sphenoid bone
	Insertion:	Medial surface of the ramus of the mandible
	Actions:	Elevate mandible. Protract mandible
	Innervation:	Trigeminal
Generate Condition – Continuously generate the missing fragments in the following muscle:		
1. Platysma	*Origin:*	Fas_ia co_erin_ s_peri_r p_rti_n of p_ctoral_s m_jor
	Insertion:	Ba_e of ma_dibl_, sk_n and l_wer part of f_ce
	Actions:	D_pr_ss ma_dibl_. Ti_hte_ n_ck fas_ia. Dr_w d_wn co_ner_ of mo_th
	Innervation:	F_cia_
2. Occipitofrontalis	*Origin:*	Gal_a apon_urot_ca
	Insertion:	Sk_n s_peri_r to e_ebr_ws and s_peri_r n_chal li_e
	Actions:	R_ise e_ebr_ws and w_ink_e fo_ehe_d. Re_ract the gal_a apon_urot_ca
	Innervation:	F_cia_
3. Medial Pterygoid	*Origin:*	M_d_al su_fa_e of lat_ral _teryg_id pl_te of sp_e_oid b_ne
	Insertion:	M_d_al su_fa_e of the r_mus of ma_dibl_
	Actions:	El_vat_ the ma_dibl_. Pr_tra_t the m_ndibl_
	Innervation:	Tri_emin_l
Test Condition – Recall from memory and fill in as much of the following information as you can:		
1. Platysma	*Origin:*	
	Insertion:	
	Actions:	
	Innervation:	
2. Occipitofrontalis	*Origin:*	
	Insertion:	
	Actions:	
	Innervation:	
3. Medial Pterygoid	*Origin:*	
	Insertion:	
	Actions:	
	Innervation:	

Table 2.

Experimental Groups and the Order in Which They Completed Each of the Components of the Studying Session

	Sequence of task completion		
Group	First	Second	Third
1	RRRR and Muscle I	RGRG and Muscle II	RTRT and Muscle III
2	RRRR and Muscle I	RGRG and Muscle III	RTRT and Muscle II
3	RRRR and Muscle II	RGRG and Muscle I	RTRT and Muscle III
4	RRRR and Muscle II	RGRG and Muscle III	RTRT and Muscle I
5	RRRR and Muscle III	RGRG and Muscle II	RTRT and Muscle I
6	RRRR and Muscle III	RGRG and Muscle I	RTRT and Muscle II
7	RGRG and Muscle I	RTRT and Muscle II	RRRR and Muscle III
8	RGRG and Muscle I	RTRT and Muscle III	RRRR and Muscle II
9	RGRG and Muscle II	RTRT and Muscle I	RRRR and Muscle III
10	RGRG and Muscle II	RTRT and Muscle III	RRRR and Muscle I
11	RGRG and Muscle III	RTRT and Muscle II	RRRR and Muscle I
12	RGRG and Muscle III	RTRT and Muscle I	RRRR and Muscle II
13	RTRT and Muscle I	RRRR and Muscle II	RGRG and Muscle III
14	RTRT and Muscle I	RRRR and Muscle III	RGRG and Muscle II
15	RTRT and Muscle II	RRRR and Muscle I	RGRG and Muscle III
16	RTRT and Muscle II	RRRR and Muscle III	RGRG and Muscle I
17	RTRT and Muscle III	RRRR and Muscle I	RGRG and Muscle II
18	RTRT and Muscle III	RRRR and Muscle II	RGRG and Muscle I

Muscle set I included the adductor hallucis, lumbricals of the foot and plantar interossei; muscle set II included the platysma, occipitofrontalis, and medial pterygoid muscles; muscle set III included the scalenes, masseter, and temporalis muscles.

completion of their second exam, the students in the experimental class section were ranked according to their current performance (i.e., score) in the course. The top eighteen performers were then randomly assigned to one of 18 experimental groups, as were the next set of eighteen students, followed by the next set, etc. The student's group assignment determined the order in which they would complete each of the components of the studying session (Table 2).

The studying session and two data collection phases (i.e., assessments) of the experiment were conducted during two predetermined class meetings. The students were informed in advance that a very small amount of course credit (5% of the total class points) depended on their attendance and full participation during both of those class meetings, but they had no prior knowledge of what they were going to be doing during either. The instructor began the first class meeting by reading a three minute script that explained: the technique associated with each of the three studying strategies, the exact procedures the students would need to carefully follow while they were using those strategies and the recall assessment they would complete immediately after they had finished the studying session. The instructor then handed each student a unique studying packet and instructed her or him to simply follow the instructions within. The purpose of the studying packet was to carefully guide the student through each predetermined phase of her or his studying session. Once all the students had received their packet, a timer was started and students began repeatedly studying their first randomly assigned muscle set using their first randomly assigned studying strategy. Every three minutes, the instructor prompted the students to stop what they were doing, turn to the next page in their packet and begin the next phase/condition of their studying strategy. Once the students had finished the 4×3 minute phases of their first studying session, they were instructed to sit quietly and/or stretch for three minutes to help clear their mind before moving on. The same studying and rest progression was repeated with the students' second randomly assigned muscle set and strategy and then again with their third.

Once the studying session was concluded, the instructor promptly collected each student's studying packet, including all the muscle set information and all the notes the student

took during the testing and generation conditions. Students then received, and immediately completed, an assessment (Immediate Assessment) that required them to recall as much of the information as they could about the nine muscles they had just studied. Exactly one week later and, again, without any prior notification, students completed the same free recall assessment a second time (Delayed Assessment). In order to reduce the likelihood of bias, the Immediate and Delayed Assessments were evaluated by a colleague who was not otherwise associated with the study and, more importantly, had no way of knowing which assessment responses corresponded to each studying strategy. That colleague held a Master's degree in kinesiology and was well versed in the subject matter. He also used a highly itemized rubric to evaluate the students' responses. According to that rubric, there were 116 possible points per assessment and each operative term (i.e., noun, verb, and adjective) was worth one point. Therefore, one point was awarded for correctly recalling each muscle's innervating nerve (e.g., masseter - trigeminal and platysma - facial), while each muscle action was worth two points because each was comprised of two directly linked terms (i.e., a specific *structure* pulled into a specific *action*). More specifically, two points were awarded for recalling both the correct structure and action (e.g., scalenes - rotate neck and temporalis - retract mandible), but no points were awarded for either an incomplete response (e.g., temporalis - _____ mandible or temporalis – elevate _____) or an incorrect action (e.g., scalenes - retract neck or scalenes - rotate mandible). Furthermore, because some of the actions and structures pertain to multiple muscles (e.g., many muscles depress something and many move the mandible), participants had to correctly match each action with each structure for each muscle (within the context of its origin and insertion) to get credit. Hence, participants received two points for indicating an action of the occipititofrontalis was retraction of the aponeurotica, but they did not receive even partial credit if they wrote, say, retraction of the mandible. As to the origin and insertion information, one point was awarded for each anatomical term that was correctly recalled (e.g., origin of the lumbricals of the foot - tendons, flexor, digitorum, and longus were each one point). In the case of muscles that had more than one origin attachment or insertion attachment, no points were awarded for anatomical terms that were grouped with the wrong attachment. For example, the studying packet described the insertion of the temporalis muscle as the *coronoid process of the mandible* and *anterior ramus of mandible*; three points were associated with the first attachment (i.e., coronoid, process and mandible) and three points with the second attachment (i.e., anterior, ramus, and mandible). If a student's answer was *coronoid ramus of the mandible and anterior process of mandible*, then she/he would have received four out of six possible points. Final note, students were not penalized for minor misspellings and no points were deducted when their answers included an incorrect anatomical term or action. However, it is important to point out that the students were far more likely to leave an answer, or a component of an answer, blank then they were to include erroneous information.

To summarize some of the key points made above, the nine muscles that were studied in this experiment were chosen because they were not otherwise covered in the Structural Kinesiology course and the students very likely had no previous experience with them. Furthermore, the students had no advanced knowledge of what they were going to be doing during the experimental class meetings, and so they had neither the basis, nor the incentive, to prepare for those meetings. Consequently, it is reasonable to assume the Immediate and Delayed Assessment specifically evaluated what the students had learned from the studying session and, by extension, how effectively each studying technique facilitated learning.

Finally, shortly after the completion of the Delayed Assessment, the students were required to indicate whether or not they: (1) had carefully followed the instructions they were given during each part of the studying session; (2) had answered every part of the Immediate and Delayed Assessments to the best of their ability; and (3) wished to participate in the study be allowing the author to use their data in the analysis. Only those students that agreed to all three of the above statements were allowed to become participants in the study.

Data Analysis

Data were analyzed using repeated measures analyses of variance (ANOVA). SPSS statistical package, version 21 (IBM Corp., Armonk, NY) was used to perform the statistical analysis. Statistical significance was set at $P < 0.05$. Assessment scores are expressed as mean percentages ± standard error.

RESULTS

A total of 69 students were enrolled in the experimental course. Of those students, 66 (96%) completed all of the experimental activities, satisfied the requirements for becoming participants and, therefore, were included in the analysis.

The mean Immediate and Delayed Assessment scores are presented in Table 3. The results were submitted to 3 × 2 repeated measures ANOVA, with studying strategy (R-R-R-R, R-G-R-G, and R-T-R-T) and assessment interval (Immediate and Delayed Assessments) as the independent variables. The analysis revealed a main effect of assessment interval $F(1, 65) = 306.15$, $P = 0.00$, $\eta^2 = 0.50$, and the scores fell an average of 75% between the Immediate and Delayed Assessments. That extent of forgetting following the one week delay, while striking, seems reasonable given that the participants had only 36 minutes (total) to learn three large sets (12 minutes each set) of difficult anatomical terms and actions. There was also a significant main effect of studying strategy $F(2, 130) = 20.84$, $P = 0.00$, $\eta^2 = 0.06$. Planned comparisons revealed the R-T-R-T strategy facilitated a higher total assessment score than both the R-R-R-R $F(1, 65) = 12.93$, $P = 0.00$, $\eta^2 = 0.17$ and R-G-R-G $F(1, 65) = 45.07$, $P = 0.00$, $\eta^2 = 0.41$ strategies, and the R-G-R-G strategy resulted in a lower total score compared to the R-R-R-R strategy $F(1, 65) = 6.38$, $P = 0.01$, $\eta^2 = 0.09$. However, these main effects were qualified by a significant interaction $F(2, 130) = 5.36$, $P = 0.01$, $\eta^2 = 0.01$, which indicated the difference in scores between the Immediate and Delayed Assessments varied by studying strategy. The decline in assessment scores following the R-G-R-G strategy was significantly less compared to the R-R-R-R $F(1, 65) = 5.20$, $P = 0.03$, $\eta^2 = 0.07$ and R-T-R-T $F(1, 65) = 10.32$, $P = 0.00$, $\eta^2 = 0.14$ strategies (24, 32, and 36 points, respectively).

In order to gain a deeper understanding of how performance varied by studying strategy, follow up tests were conducted using three sets of repeated measures ANOVAs. First,

Table 3.

Mean Recall Scores As a Function of Studying Strategy

Assessment	R-R-R-R studied questions	R-G-R-G studied questions	R-T-R-T studied questions	Assessment mean
Immediate	40.01 ± 2.92	32.06 ± 2.75	49.96 ± 3.24	40.79 ± 2.24
Delayed	8.56 ± 0.90	8.03 ± 1.00	14.46 ± 1.66	10.39 ± 0.92
Strategy mean	24.29 ± 1.62	20.04 ± 1.70	32.21 ± 2.16	

Scores are expressed as mean percentages ± standard error.

on the Immediate Assessment, the R-R-R-R strategy facilitated significantly greater recall scores $F(1, 65) = 6.52$, $P = 0.01$, $\eta^2 = 0.09$ than the R-G-R-G strategy, but the R-T-R-T strategy lead to the highest scores $F(1, 65) = 7.12$, $P = 0.01$, $\eta^2 = 0.09$ and $F(1, 65) = 28.60$, $P = 0.00$, $\eta^2 = 0.31$, respectively. On the Delayed Assessment, there was no longer a statistical difference $F(1, 65) = 0.22$, $P = 0.64$ between the R-R-R-R and R-G-R-G strategies, but the R-T-R-T strategy still produced the greatest recall scores $F(1, 65) = 11.29$, $P = 0.00$, $\eta^2 = 0.15$, and $F(1, 65) = 24.23$, $P = 0.00$, $\eta^2 = 0.27$, respectively. Thus, the R-T-R-T condition yielded consistently superior retention throughout the assessment period. This relationship was further confirmed by a descriptive analysis of forgetting relative to initial recall, using the procedure followed by Roediger and Karpicke (2006): (Immediate Assessment score – Delayed Assessment score)/Immediate Assessment score (Fig. 1).

DISCUSSION

The results of the current study comparing university-level students' learning of difficult anatomical concepts showed that, on both Immediate and Delayed assessments, the R-T-R-T strategy was superior to a R-R-R-R and a R-G-R-G strategy. The pertinent effect sizes ranged from medium to large, which lends credence to the conclusion that the R-T-R-T condition was an effective strategy for university-level participants to employ. The results point to the conclusion that the variant of the classic study-test condition yielded superior retention of anatomical concepts for university-level students. Thus, the act of retrieving information during the learning phase, as opposed to simply reviewing materials repeatedly, results in better memory for anatomical concepts. And, the most challenging, and unconstrained, retrieval task performed during the learning acquisition phase enhanced retention of anatomical concepts both immediately and in the long term.

Some interesting patterns emerged from this investigation that require additional discussion. As in Dobson and Linderholm (2015), the variant of the traditional study-test condition yielded superior performance at both the immediate and delayed assessment points. This is counter to the majority of the literature, although there are exceptions (e.g., Meyer and Logan, 2013), where typically the study-test condition yields inferior performance at the immediate delay but superior performance at the longer delay. The methodology used previously by Dobson and Linderholm (2015) did not constrain how much time was spent learning in each strategy condition, which was thought to possibly account for the unique pattern in the former study. That is, it could have been that students capitalized on the lack of time constraints when in the study-test condition and spent more time testing themselves. But in the current study, time constraints were carefully controlled across each of the three strategy conditions, allowing only three minutes per study session. Thus, there appears to be some consistency across two investigations using difficult anatomy and physiology information that self-testing is effective both in the immediate and long term for university-level students. This is encouraging as students who experience the positive reward of immediate learning may be more likely to engage in that strategy consistently over time (Kornell and Bjork, 2007).

An explanation for why the testing effect was found even in the immediate delay condition should be considered. What this study has in common with other studies that have found similar benefits of testing even in the immediate delay period (e.g., Meyer and Logan, 2013; Dobson and Linderholm, 2015) is that these are all situations where research participants were studying materials that were relevant to their

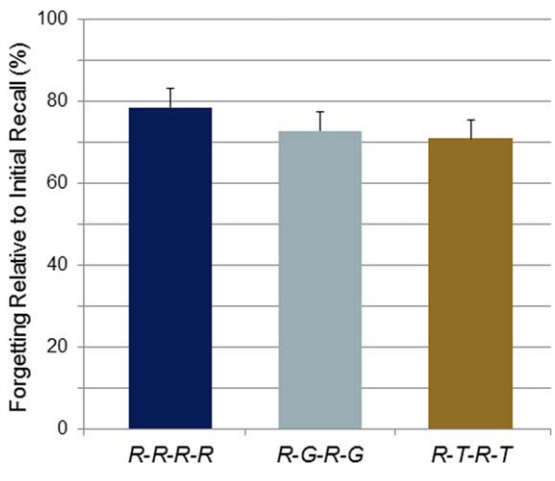

Figure 1.

Forgetting relative to initial recall as a function of studying strategy. Forgetting was determined according to the following equation: [(Immediate Assessment score – Delayed Assessment score)/Immediate Assessment score] × 100. The R-R-R-R, R-T-R-T and R-G-R-G abbreviations refer to the reading-based, testing-based, and generation-based studying strategies, respectively. Error bars represent the positive standard error of the means.

experiences, so they were familiar with the content of the to-be-learned items. For example, laboratory-based studies using less familiar materials (e.g., Roediger and Karpicke, 2006; Karpicke and Blunt, 2011) do not show an advantage of testing at immediate delays but more applied studies using relevant, possibly familiar materials to the research participants do (e.g., Meyer and Logan, 2013; Dobson and Linderholm, 2015). Thus, future investigations should systematically manipulate the relevance or familiarity of the to-be-learned items to ascertain whether or not it is a factor that influences at what time delays the benefits of testing can be observed.

Another unique pattern found in this study was that generation was not, at the very least, a more effective strategy than the passive R-R-R-R strategy. This is counter to the majority of the literature (e.g., Hirshman and Bjork, 1988; Richland et al., 2005) and we present several explanations for this finding. First, the word fragment task was not well executed in terms of using a standardized method for developing materials as suggested by Koopman et al. (2013). Because an item analysis was not done prior to using these word fragment tasks, it may not be a reliable measure for distinguishing between participants' performance. Another related possibility is that the word fragment task did not force participants to retrieve anatomical information from memory. Perhaps some participants were able to guess missing letters without relying on memory but simply guessed the correct answer based on logic. So, in other words, performance on the word fragment task was not necessarily a reflection of how well participants retrieved information from memory but more of an indicator of how good the participant was in guessing terms correctly or using logic to fill in letters. Finally, it could be that the word fragment task itself focused attention on superficial qualities of the to-be-learned items such as spelling and not meaning. Focusing on superficial aspects of the items could have inhibited elaborative processing, which would lessen retention. Given these potential methodological flaws, it is perhaps premature to claim that the word fragment task may not be useful for retention of anatomy terms. Further investigations should construct methodologically sound word fragment tasks to test the possibility, albeit remote, that generation could be a reasonable learning strategy to use in anatomy courses.

Limitations

Despite the optimistic findings of the current study for enhancing the retention of anatomical information, it has some limitations. First, the successful study strategy, that is, self-testing, was implemented under the watchful eyes of researchers in somewhat of an artificial environment. It is unknown whether or not students would actually employ this strategy when studying for their courses independently. Although there is evidence that students are capable of using this strategy on their own once they understand the benefit of self-testing (e.g., Dobson and Linderholm, 2015), this point was not pursued in the current study. Likewise, the time delay used in this investigation spanned one week. Ideally, we hope that students retain information for longer periods of time than one week and, at a minimum, until final exams. From the current study design, it is impossible to know how long the anatomy information would be retained over the course of a semester or beyond. Even though the study has limitations, there are still useful tips that an anatomy instructor can garner from the results. First, instructors could discuss explicitly the strategy and how it has been shown to be a superior learning strategy in this and other studies. Showing data and results from this and other studies may be motivating to students who appreciate evidence-based approaches to learning. Second, instructors could allow time at the end of each lecture/class period for students to engage in free recall. This would help students retain information from that class session. After the strategy is modeled within the constraints of classroom meetings, students can be encouraged to use the same technique after reading textbooks/course materials independently. Third, instructors could occasionally give feedback to students about their ability to retain lecture information to reinforce the strategy. That is, after students are asked to free recall at the end of the lecture, instructors could give them feedback about the accuracy and completeness of their memories. Feedback is a valuable tool for giving students a sense of how accurate their self-assessments of learning are and whether or not they need to take corrective action. The three instructional recommendations could be useful to any anatomy instructor and a logical extension of current results, despite the stated limitations of the current study.

CONCLUSIONS

To conclude, based on the results of this investigation and another investigation using similar materials and participants (Dobson and Linderholm, 2015) we propose that at least some anatomy materials are best learned by requiring university-level students to actively free recall to-be-learned information during the initial learning phase. Anatomy students should be urged to first review materials and then put them aside to allow for time for retrieval attempts. This advice applies to a variety of assessment formats (e.g., multiple choice, free recall) so should be effective for a host of university-level testing situations in anatomy courses.

NOTES ON CONTRIBUTORS

JOHN DOBSON, Ph.D., is an assistant professor in the Department of Health and Kinesiology, College of Health and Human Sciences, Georgia Southern University, Statesboro, Georgia. He teaches anatomy and physiology classes to undergraduate and graduate kinesiology and health science students.

TRACY LINDERHOLM, Ph.D., is a professor of educational psychology and Associate Dean of Graduate Education and Research, College of Education, Georgia Southern University, Statesboro, Georgia. Her research has focused on the cognitive-psychological processes involved in learning and comprehending text information.

LITERATURE CITED

Bjork RA. 1994. Memory and metamemory considerations in the training of human beings. In: Metcalfe JA, Shimamura AP (Editors). Metacognition: Knowing about Knowing. 1st Ed. Cambridge, MA: MIT Press. p 185–205.

Bjork RA, Bjork EL. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In: Gernsbacher MA, Pew RW, Hough LM, Pomerantz JR (Editors). Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society. 1st Ed. New York, NY: Worth Publishers. p 56–64.

Bjork RA, Dunlosky J, Kornell N. 2013. Self-regulated learning: Beliefs, techniques, and illusions. Annu Rev Psychol 64:417–444.

Butler AC, Roediger HL III. 2007. Testing improves long term retention in a simulated classroom setting. Eur J Cogn Psychol 19:514–527.

Carpenter SK, DeLosh EL. 2006. Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. Mem Cognit 34:268–276.

Dobson JL, Linderholm T. 2015. Self-testing promotes superior retention of anatomy and physiology information. Adv Health Sci Educ Theory Pract 20: 149–161.

Einstein GO, Mullet HG, Harrison TL. 2012. The testing effect: Illustrating fundamental concept and changing study strategies. Teach Psychol 39:190–193.

Glover JA. 1989. The "testing" phenomenon: Not gone but nearly forgotten. J Educ Psychol 81:392–399.

Hinze SR, Wiley J. 2011. Testing the limits of testing effects using completion tests. Memory 19:290–304.

Hirshman E, Bjork RA. 1988. The generation effect: Support for a two-factor theory. J Exp Psychol Learn Mem Cognit 14:484–494.

Jönsson FU, Hedner M, Olsson MJ. 2012. The testing effect as a function of explicit testing instructions and judgments of learning. Exp Psychol 59:251–257.

Karpicke JD, Blunt JR. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. Science 331:772–775.

Koopman, J, Howe M, Johnson RE, Tan JA, Chang CH. 2013. A framework for developing word fragment completing tasks. Hum Resource Manag Rev 23:242–253.

Kornell N, Bjork RA. 2007. The promise and perils of self-regulated study. Psychon Bull Rev 14:219–224.

Larsen DP, Butler AC, Roediger HL 3rd. 2009. Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. Med Educ 43:1174–1181.

Larsen DP, Butler AC, Lawson AL, Roediger HL 3rd. 2013. 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 18:409–425.

Larsen DP, Butler AC, Roediger HL 3rd. 2013. Comparative effects of test-enhanced learning and self-explanation on long-term retention. Med Educ 47: 674–682.

Lipowski SL, Pyc MA, Dunlosky J, Rawson KA. 2014. Establishing and explaining the testing effect in free recall for young children. Dev Psychol 50: 994–1000.

Logan JM, Thompson AJ, Marshak DW. 2011. Testing to enhance retention in human anatomy. Anat Sci Educ 4:243–248.

Meyer AN, Logan JM. 2013. Taking the testing effect beyond college freshman: Benefits for lifelong learning. Psychol Aging 28:142–147.

Pyc MA, Rawson KA. 2010. Why testing improves memory: Mediator effectiveness hypothesis. Science 330:335.

Richland LE, Bjork RA, Finley JR, Linn MC. 2005. Linking cognitive science to education: Generation and interleaving effects. In: Bara BG, Barsalou L, Bucciarelli M (Editors). In: Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society (CogSci2005). Stresa, Italy, 2005 July 21–23. p 555–583. Cognitive Science Society, Wheat Ridge, CO.

Roediger HL III, Karpicke JD. 2006. The power of testing memory: Basic research and implications for educational practice. Perspect Psychol Sci 1:181–210.

Zaromb FM, Roediger HL 3rd. 2010. The testing effect in free recall is associated with enhanced organization processes. Mem Cognit 38:995–1008.