Adv in Health Sci Educ (2015) 20:149–161
DOI 10.1007/s10459-014-9514-8

Self-testing promotes superior retention of anatomy and physiology information

John L. Dobson · Tracy Linderholm

Received: 19 December 2013 / Accepted: 5 May 2014 / Published online: 17 May 2014

Abstract The testing effect shows that learning is enhanced by the act of recalling information after exposure. Although the testing effect is among the most robust findings in cognitive science, much of its empirical support is from laboratory studies and it has been applied as a strategy for enhancing learning in the classroom in a limited fashion. The purpose of this investigation was to replicate the testing effect in a university anatomy and physiology course and to extend the applicability of it to independent student study. Students repeatedly studied three sets of passages that described structures and concepts pertaining to (1) cardiac electrophysiology, (2) ventilation and (3) endocrinology. Each student was randomly assigned to study one of those three passage sets by reading it three consecutive times (R–R–R), another by reading and then rereading it while taking notes (R–R + N) and the third by reading it, recalling as much as possible (i.e., self-testing) and then rereading it (R–T–R). Retention assessed after 1 week was significantly greater following R–T–R (53.95 ± 1.72) compared to R–R–R (48.04 ± 1.83) and R–R + N (48.31 ± 1.78). Evidence is also presented that suggests students benefited from instructions to self-test when preparing for exams on their own. The testing effect, then, can be generalized to real-life settings such as university anatomy and physiology courses and to independent study situations.

Keywords Testing · Self-testing · Retrieval practice · Learning · Anatomy and physiology

J. L. Dobson (✉)
Department of Health and Kinesiology, Georgia Southern University, P.O. Box 8076, Statesboro, GA 30460, USA
e-mail: jdobson@georgiasouthern.edu

T. Linderholm
Department of Curriculum, Foundations, and Reading, Georgia Southern University, Statesboro, GA, USA

Introduction

There is great concern among politicians and educators alike that the United States does not train science majors who are well-prepared go into career fields such as the health professions, medicine and engineering (National Academy of Sciences 2007; National Commission on Mathematics and Science Teaching for the 21st Century 2000). One approach to better prepare science students is to equip them with learning techniques that are optimal for the complex information they will encounter in university science courses. To pursue this approach, it is necessary to draw from well-established memory and learning techniques from the field of cognitive science. Cognitive science is the empirical study of the human mind and brain, and experiments are typically conducted in controlled laboratory settings. Cognitive science experiments have shown one technique to be particularly useful for long-term retention—it is generally known as the testing effect or retrieval practice (e.g., Karpicke and Blunt 2011; Roediger and Karpicke 2006). The testing effect refers to the robust finding that individuals who actively engage in self-testing as a learning strategy show long-term retention benefits compared to those who simply review or restudy the same materials. The testing effect is robust across age groups ranging from elementary-aged school children to middle aged and/or older adults (e.g., Lipowski et al. 2014; Meyer and Logan 2013); is present with different forms of assessment (Roediger et al. 2011) including neuroimaging techniques (Wing et al. 2013); and has been detected in a variety of content areas including psychology and medicine (Einstein et al. 2012; Larsen et al. 2008; Logan et al. 2011). With this in mind, the first objective of the current study is to replicate the testing effect in a university anatomy and physiology course. The second objective is to determine if anatomy and physiology students can appreciate the benefits of the testing technique upon exposure to it and employ it on their own, thereby extending the generalizability of the testing effect well beyond an immediate classroom intervention.

The testing effect has been shown to enhance memory and retention of concepts, across different content areas, particularly after a period of delay as long as 1–2 weeks (Roediger and Karpicke 2006). The typical protocol followed in investigations of the testing effect involves a particular sequence: participants review the materials, engage in a recall attempt, and then their retention is measured at varying delay periods (e.g., 5 min, 1 day, 1 week). The act of recalling information is a learning episode in and of itself that strengthens the memory much more than two simple exposures to the same learning stimuli alone where participants may be asked to review the materials twice before their retention is measured. In other words, testing is viewed as a learning strategy rather than as only a final measure of one's knowledge or level of understanding (Roediger and Karpicke 2006). Studies often involve comparisons between participants learning new materials in a condition where materials are studied and then re-studied, that is, a "study–study condition", compared to a condition where new materials are studied and then recalled, that is, a "study-test condition". Results across multiple experiments are consistent and show that the study-test condition yields superior retention of materials, particularly after a time delay such as 2 days or even up to 1 week after materials were first encountered (e.g., Karpicke and Blunt 2011; Roediger and Karpicke 2006). Attempting to recall materials rather than passively reviewing or restudying them appears to be a more cognitively demanding task and creates what cognitive scientists call "desirable difficulties" during the learning phase (Bjork 1994; Bjork and Bjork 2011).

Desirable difficulties is a theoretical tenet used to explain why some strategies, such as the testing effect, interleaved stimulus presentation, and spaced practice, do not necessarily

improve short-term retention but are superior to other strategies in the long-term (Bjork 1994; Roediger and Karpicke 2006). The concept of desirable difficulties suggests that learning strategies that are more active and cognitively demanding during the encoding or learning phase, and require additional attention, may disrupt short-term recall but aid in building the necessary networks in memory to enhance long-term recall. For example, learning materials in an interleaved sequence is harder, and more cognitively demanding than learning materials in a blocked presentation format. Interestingly, learning in a blocked format may lead to short-term retention benefits but, in the long-term, learning in an interleaved presentation format is superior. Similarly, "cramming for an exam", also known as massed practice, may allow a student to earn an "A" on an exam taken the next day but this learning strategy will not be helpful for recalling the information covered on the exam in years to come. However, spaced practice, learning the items slowly over time and over several study sessions, leads to superior long-term retention because it creates desirable difficulties (e.g., Bjork 1994; Bjork and Bjork 2011). Thus, from a theoretical standpoint, the testing effect works because it increases cognitive demand during the encoding phase and this produces better retention. Compared to passively reviewing the material, that requires minimal cognitive effort, the act of testing requires more work on the part of the learner. The notion that creating desirable difficulties at encoding is superior for long-term retention is well supported by behavioral evidence as well as neuroimaging evidence (Wing et al. 2013).

Desirable difficulties in the context of testing effect investigations are most commonly conducted in a laboratory setting (Karpicke and Blunt 2011; Roediger and Karpicke 2006) and make use of fairly simple materials. Having a laboratory setting with fairly simple materials is important for controlling extraneous variables that may impact results and showing the robust nature of an effect. An even stronger indicator of the testing effect would be if the pattern holds up in real life situations and on more complex materials such as those found in anatomy and physiology courses. Several researchers have been successful in showing the testing effect in different contexts (e.g., Einstein et al. 2012; Larsen et al. 2009; Logan et al. 2011; Lyle and Crawford 2011; McDaniel et al. 2011; Roediger et al. 2011). For example, Einstein et al. (2012) showed the testing effect in the context of an undergraduate psychology laboratory course. Students were presented with data from their own classroom experiment that provided quantitative evidence of the benefits of testing. Students themselves participated in the experiment and the to-be-learned materials were presented in either a study–study or a study-test condition. The results showed that the students in the study-test condition performed better on a memory test of the to-be-learned items than those in the study–study condition. Thus, it is likely that the testing effect can be applied to learning in real-life situations, such as university classrooms, but the complexity of a variety of situations warrants replication, as is one of the objectives of the current study.

An interesting common result from some of the testing effect investigations, both laboratory based studies and real-life applications, is the lack of knowledge that fairly sophisticated university students possess regarding what learning strategies are most beneficial for comprehending and retaining university-level material (Bjork et al. 2013; Karpicke and Blunt 2011; Einstein et al. 2012; Kornell and Bjork 2007). In general, university students seem to be unaware of the benefits of testing as a learning strategy and prefer other, perhaps less effortful, strategies such as rereading to-be-learned material or taking notes (Karpicke and Roediger 2008; Kornell and Bjork 2007; Kornell and Bjork 2009). A ray of hope relevant to the current investigation is that, in at least in a few cases, college-level participants show some recognition of the benefits of the strategy (Einstein

et al. 2012; Tullis et al. 2013). Specifically, the results from Einstein et al. (2012) showed that university students eventually adopted a testing strategy after being allowed to inspect their own data from a testing effect study they themselves participated in and after seeing the benefits of employing the strategy at exam time. So providing university students with concrete evidence of strategy effectiveness may be a way to encourage them to use beneficial strategies independently, which is a second, and perhaps most innovative, objective of the current study.

To reiterate, the current study had two objectives relevant to the literature reviewed: (1) to provide converging evidence for the testing effect in a university anatomy and physiology classroom context using materials relevant to students' majors, a real-life educational context; and (2) to determine if students would continue to benefit from learning the testing strategy when preparing independently for course exams. To meet both objectives, two phases to the investigation were developed. In Phase I, university student participants were asked to study three anatomy and physiology texts using a rereading strategy, a note-taking strategy, and a retrieval practice strategy. The three conditions were: (1) Read, Read, Read (R–R–R); (2) Read, Take Notes + Read (R–R + N); and (3) Read, Free Recall/Test, Read (R–T–R). Participants in each condition were encouraged verbally and in writing to spend 15–20 min on each portion of the learning sequence to equalize time spent in each condition. Upon experiencing each of the three conditions, participants were immediately tested and then again 1 week later. In Phase II, the results of Phase I were shared with students in the course in an attempt to encourage them to continue using the best strategy to prepare independently for course exams. Course exam scores from students in the course were compared to course exam scores of students (in different sections of the course) who did not receive the same information to determine if students were able to employ the strategy on their own.

Methods

Participants

All experimental procedures were approved by the University's Institutional Review Board. The participants were recruited from three sections of an Anatomy and Physiology I (A&P I) course at a regional US university. The experiment was broken up into two different phases (Phases1 and 2); Phase 1 participants were all recruited from a Fall 2012 section, whereas students who participated in Phase 2 were recruited from Spring 2012 and Fall 2011 sections (Control Groups A and B, respectively). The typical A&P I student was either a second or third year undergraduate student and was an allied health or similar major (e.g., pre-physical therapy, pre-nursing, exercise science, pre-medicine, etc.).

Phase 1 treatment

The students that completed this Phase all repeatedly studied three different anatomy and physiology passages that provided overviews of: 1. cardiac electrophysiology (621 words), 2. endocrinology (644 words) and 3. ventilation (650 words). These three topics were chosen specifically because they were not part of the material covered in the A&P I course, but they were covered in subsequent courses. The A&P instructor identified twenty to thirty key definitions and concepts pertaining to each of those three A&P II topics and used them to write each of the three passages for Phase 1 (see sample text, Table 1). Students

used each of the following strategies to study the experimental passages but used only one strategy per passage. One strategy required students to carefully read a passage three consecutive times (R–R–R strategy). Another strategy required students to carefully read a passage and then immediately reread it while also writing down any notes they thought important (R–R + N strategy). The third studying strategy required students to carefully read a passage, then immediately test themselves on (i.e., freely recall) that information and, finally, reread the passage (R–T–R). During the testing portion of the R–T–R strategy, the students could not see the information from the pertinent passage and they were instructed to write down as many of the important definitions and concepts as they could on a provided blank sheet of paper. Finally, students were instructed to spend no more than 5–7 min conducting each portion of each studying strategy (e.g., taking notes during the note taking portion of the R–R + N strategy or recalling information during the testing portion of the R–T–R strategy).

The experimental passages and strategies described above were administered both in a sequential order and during only one studying session. Therefore, in an effort to remove any potential effects associated with studying strategy order, the students used the strategies in an order that was both randomized and specified by a group assignment.

Phase 1 procedure

All A&P I students were required to complete five exams throughout the semester. After the second course exam had been graded, the Phase 1 students were matched according to their exams scores and then randomly assigned to a studying strategy order group.

The studying session and two assessments that constituted this Phase were conducted during two pre-determined class meetings. During the first of those meetings, the instructor began the class by thoroughly describing: the technique associated with each of the three studying strategies, the procedures the students would use throughout the subsequent studying session and the Immediate Assessment they would complete after studying. The instructor included numerous calls for the students to carefully read and precisely follow the instructions they were given, so that they were sure to properly perform each studying strategy. The instructor then handed each student a unique studying packet that would serve as a guide to help her or him properly complete all components of the studying phase. Students then began the studying session by repeatedly studying their first randomly assigned passage using their first randomly assigned studying strategy. The studying packet instructed students to spend both no more than 5–7 min on each portion of a studying strategy and no more than a total of 15–20 min studying an individual passage. The instructor periodically made an announcement throughout the studying session to encourage the students to stay on track and not spend too much time on any one portion or passage.

Directly after they completed their studying session, the students returned all of the pertinent materials, including: the three anatomy and physiology text passages they studied, the notes they took with R–R + N strategy and the information they recalled and recorded during the R–T–R strategy. The students then received, and immediately completed, a thirty multiple choice question assessment (Immediate Assessment). Ten of the questions pertained to the key definitions and concepts from the cardiac electrophysiology passage, ten pertained to the endocrinology passage and ten pertained to the ventilation passage (see sample questions, Table 2). Note, the sets of ten questions corresponding to the material that students learned via R–R–R, R–R + N and R–T–R strategies will be abbreviated 10_{R-R-R}, 10_{R-R+N} and 10_{R-T-R}, respectively. During the class meeting that

Table 1 A portion of the ventilation passage that participants repeatedly studied

One of the components of respiration is **ventilation**, which specifically refers to the mechanical action of moving gasses into and out of the lungs. Gasses are pulled from the atmosphere into the lungs during **inspiration**; whereas gasses are pushed from the lungs into the atmosphere during **expiration**. Both inspiration and expiration are accomplished via a **pressure gradient**, which means that gasses move by bulk flow from an area of higher pressure to an area of lower pressure
The two gas pressures that are important to ventilation are the pressure of gas: (1) in the atmosphere (**atmospheric pressure**) and (2) inside the lungs (**intrapulmonary pressure**). The closer a reference point is to sea-level, the taller the column of atmosphere, and the greater the atmospheric pressure, will be at that point. Consequently, the atmospheric pressure is relatively high at sea-level (e.g., 760 mmHg) and it progressively decreases as one ascends higher into altitude (e.g., is ~630 mmHG at 5,000 ft, ~520 mmHG at 10,000 ft, and so on)
The second important gas pressure is the intrapulmonary pressure, which we manipulate so that it cyclically rises above the atmospheric pressure, then falls below it, then rises above it, etc. When we increase the intrapulmonary pressure above the atmospheric, the resulting pressure gradient pushes gas out of the lungs and into the atmosphere. When we decrease the intrapulmonary pressure below the atmospheric, the gradient pulls gas from the atmosphere into the lungs
Our ability to manipulate intrapulmonary pressure is possible because of **Boyle's Law**, which states that the pressure of a gas is inversely proportional to its volume. The volume of gas in the lungs is altered by the action of the ventilatory muscles, most notably, the **diaphragm**. When the diaphragm contracts, it causes the chest cavity to expand, which, in turn, increases the volume of the lungs. As the lung volume increases, the intrapulmonary gas pressure decreases and causes inhalation
Immediately following inspiration, the diaphragm relaxes and the lungs recoil to a due mostly to the elastic pull of the tissues in the chest cavity. As the lung volume decreases, the intrapulmonary gas pressure increases and causes exhalation

The portion above contains 355 words, which is roughly 55 % of the 650 word passage that was used in the study

occurred 1 week later, the students completed the same thirty question assessment a second time (Delayed Assessment). A 1 week retention interval was selected because it has been used by numerous similar investigations that examined long-term retention following a brief period of studying (Karpicke and Blunt 2011; McDaniel et al. 2009).

It is important to point out that the students had no prior knowledge that they would be completing either the studying session involving the three study strategy conditions or that they would be assessed on their learning. Since the students also had little or no previous experience with the experimental information, it is likely that the Immediate and Delayed Assessments did specifically evaluate what the students had learned and retained from the studying session and, by extension, how effectively each studying technique facilitated learning.

Finally, 1 week after the students completed the Delayed Assessment, they were required to complete a questionnaire on which they had to indicate: 1. did they carefully follow the instructions they were given during each part of the studying session; 2. did they answer all questions on the Immediate and Delayed Assessments to the best of their ability; and 3. did they want to participate in the study be allowing the author to use their data in the analysis. Only those students that answered "yes" to all three of the above questions became participants in Phase 1 of the study.

Phase 2 treatment

Shortly after finishing the procedures described above, the A&P I instructor presented the Immediate and Delayed Assessment results to the Phase 1 students and used those findings

Table 2 Selected questions from the immediate and delayed assessments

Among the structures listed below, _____ is/are included in the respiratory zone and _____ is/are included in the conducting zone. 1 - alveoli 2 - bronchioles 3 - mouth and nose 4 - respiratory bronchioles 5 – trachea A. 1: 2, 3 and 5 B. 1: 2, 3, 4 and 5 C. 1 and 4: 2, 3 and 5 D. 2, 3 and 5: 1 and 4 E. none of the above
The higher one ascends above sea-level, the _____ pressure will be. A. lower the atmospheric B. higher the atmospheric C. higher the intrapulmonary D. answers A and C. E. mark this answer if there is no relationship between altitude and gas pressure.
When the diaphragm contracts, it causes lung volume to _____, which, in turn, causes intrapulmonary pressure to _____. A. increase: increase B. increase: decrease C. decrease: increase D. decrease: decrease E. there is no relationship between diaphragm contraction, lung volume and intrapulmonary pressure
Boyle's law states that: A. gases move from areas of high pressure to areas of low pressure. B. the pressure of a gas is inversely proportional to its volume. C. the pressure of a gas increases as the volume increases. D. the atmospheric pressure depends on a column of gas in the atmosphere. E. none of the above.

to encourage them to incorporate the superior strategy, that is, the testing-based studying technique (e.g., R–T–R), into their preparation for the remaining three course exams. The instructor also showed the students how they could develop such a studying technique by simply modifying the R–T–R strategy they had already practiced. During the class meetings that occurred immediately prior to the third, fourth and fifth course exams, the instructor once again encouraged the students to use a testing-based strategy while preparing for their exams.

Phase 2 procedure

The purpose of Phase 2 of the experiment was to determine if the evidence collected from Phase 1, the resulting class discussion and subsequent encouragement from the instructor (i.e., if the incorporation of a testing-based studying strategy), would help A&P I students better internalize the course material and increase performance on class exams. All the students that had been encouraged to use a testing-based studying strategy when preparing for course exams 3–5, and who had also agreed to participate in the experiment, were placed into a group for Phase 2 (Testing Group). By contrast, A&P I students that had not been encouraged to use a testing-based studying strategy in the course, and most were likely not even familiar with such a strategy, served as controls (Control Groups A and B).

These latter groups of participants had previously been recruited from two earlier sections of A&P I. It is important to emphasize that the information covered, the class materials (e.g., lecture slides and class notes) and the exam questions used in all three A&P I sections identified above were identical. That is, the only meaningful difference between the experimental groups was that those in the Testing Group completed the activities of Phase 1 and were subsequently encouraged to incorporate testing into their course studying habits.

Data analysis

All assessment and course exam scores were analyzed using analyses of variance (ANOVAs). Kuder-Richardson (KR-20) tests were used to evaluate the reliability of the Immediate/Delayed Assessment and all five course exams. Statistical significance was set at $P < 0.05$. Data are expressed as mean percentages ± standard error.

Results

A total of 147 students were enrolled in the A&P I class that completed Phase 1; 22 of those students either failed to complete all the required components or did not consent to having their data used in this study, whereas the remaining 125 students all became Phase 1 participants and were included in the analysis. Of those 125 participants, 120 completed all five A&P I course exams and were therefore included in the Phase 2 Testing Group. The other 248 Phase 2 participants comprised the Control Group (114 and 134 in Control Groups A and B, respectively).

Phase 1 assessment comparisons

The instrument used to assess immediate and delayed recall had a reliability coefficient (KR-20) of 0.77. The mean Immediate and Delayed Assessment scores corresponding to each of the three studying strategies are presented in Table 3. A 2 (Immediate vs. Delayed Assessments) × 3 ($10_{R–R–R}$, $10_{R–R+N}$ and $10_{R–T–R}$ scores) repeated measures ANOVA revealed a main effect of assessment interval $F(1, 123) = 160.09$, $P = 0.00$, $\eta_p^2 = 0.57$. There was also a significant main effect of studying strategy $F(2, 246) = 3.60$, $P = 0.03$, $\eta_p^2 = 0.05$. Comparisons revealed that there was no difference in scores between either the R–R–R and R–R + N strategies $P = 0.73$ or the R–R + N and R–T–R strategies $P = 0.39$, but there was a significant difference between the R–R–R and R–T–R strategies $P = 0.03$.

However, the main effects were qualified by a significant interaction $F(2, 246) = 3.89$, $P = 0.02$, $\eta_p^2 = 0.03$ between studying strategy and assessment interval. Planned comparisons revealed that a significant amount of forgetting occurred between the Immediate and Delayed Assessments following all three studying strategies $F(1, 123) = 69.19$, $P = 0.00$, $\eta_p^2 = 0.36$; $F(1, 123) = 169.02$, $P = 0.00$, $\eta_p^2 = 0.58$; and $F(1, 123) = 28.34$, $P = 0.00$, $\eta_p^2 = 0.20$ for the R–R–R, R–R + N and R–T–R strategies, respectively. Because the decrease in scores between the Immediate and Delayed Assessments was expected, follow up tests were used to determine how both immediate and delayed performance varied by studying strategy. The Immediate Assessment comparisons revealed there was no statistical difference between the R–R + N and R–T–R strategies $F(1, 123) = 0.01$, $P = 0.93$, but both resulted in significantly better performance than R–R–R

Table 3 Results of the immediate and delayed assessments from Phase 1

Assessment	R–R–R studied questions	R–R + N studied questions	R–T–R studied questions	Assessment mean
Immediate	59.76 ± 1.58	63.87 ± 1.69	64.03 ± 1.81	62.55 ± 1.33
Delayed	48.06 ± 1.83	48.31 ± 1.78	53.95 ± 1.72	50.11 ± 1.09
Strategy Mean	53.91 ± 1.56	56.09 ± 1.63	58.99 ± 1.49	

strategy F (1, 123) = 5.66, $P = 0.02$, $\eta_p^2 = 0.04$ and F (1, 123) = 4.64, $P = 0.03$, $\eta_p^2 = 0.04$, respectively. On the Delayed Assessment, there was no difference in scores associated with the R–R–R and R–R + N strategies F (1, 123) = 0.01, $P = 0.92$, but both were significantly lower than that with the R–T–R strategy F (1, 123) = 6.02, $P = 0.02$, $\eta_p^2 = 0.05$ and F (1, 123) = 4.86, $P = 0.03$, $\eta_p^2 = 0.04$, respectively. Thus, the R–T–R strategy facilitated the greatest retention of the experimental concepts following a 1-week delay.

Phase 2 course exam comparisons

The KR-20 reliability coefficients for course exams 1, 2, 3, 4 and 5 were 0.84, 0.70, 0.82, 0.78, 0.84 and 0.82, respectively. A 2 × 5 mixed factorial ANOVA was used to compare differences between Group Type (Testing Group; Control Group) across the five course exams. There was a main effect of exam number F (4, 363) = 64.47, $P = 0.00$, $\eta_p^2 = 0.2$, and planned difference contrasts indicated that exam scores generally improved significantly throughout the semester, F (1, 366) = 35.87, $P = 0.00$, $\eta_p^2 = 0.09$; F (1, 366) = 155.19, $P = 0.00$, $\eta_p^2 = 0.30$; F (1, 366) = 9.02, $P = 0.00$, $\eta_p^2 = 0.02$; and F (1, 366) = 82.22, $P = 0.00$, $\eta_p^2 = 0.18$, for exams 2, 3, 4 and 5, respectively. However, the comparison between the Testing and Control Group also revealed a significant interaction F (4, 363) = 15.07, $P = 0.00$, $\eta_p^2 = 0.04$, as well as a significant main effect of Group Type, F (1, 366) = 10.11, $P = 0.00$, $\eta_p^2 = 0.03$. Follow up tests revealed there were no statistical differences between groups on course exams 1 and 2, F (1, 366) = 0.80, $P = 0.37$; F (1, 366) = 1.03, $P = 0.32$, respectively, but those in the Testing Group had significantly higher scores on exams 3, 4, and 5 than the Control Group, F (1, 366) = 34.37, $P = 0.00$, $\eta_p^2 = 0.09$; F (1, 366) = 4.12, $P = 0.04$, $\eta_p^2 = 0.01$; F (1, 366) = 27.99, $P = 0.00$, $\eta_p^2 = 0.07$, respectively. That is, the two groups performed the same before the Testing Group was encouraged to use a testing strategy to prepare for the course exams, but, following the onset of strategy instruction and encouragement, they outscored the Control Group on the remaining three exams (Fig. 1).

Discussion

The objectives of the current study were twofold: (1) To determine whether the testing effect could be replicated in a university anatomy and physiology course; and (2) To examine whether or not students in the anatomy and physiology course would be able to recognize and then generalize the benefits of testing by employing the strategy when studying independently. With regard to the first objective, the results revealed that the testing effect was indeed present in this applied setting using fairly complex science

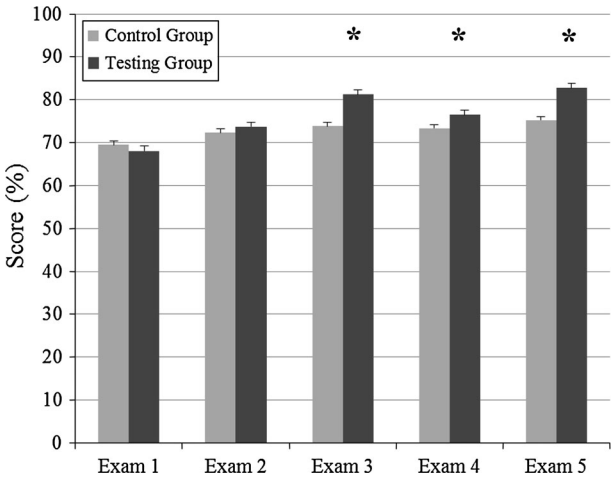

Fig. 1 Course exam scores as a function of group in Phase 2. *Indicates a significant difference on an exam score. *Error bars* represent the positive standard error of the means. Testing Group Exams 1–2 were pre-treatment and Exams 3–5 were post-treatment

materials. Specifically, students learning new anatomy and physiology text materials in the R–T–R studying condition showed superior performance compared to when learning in the R–R–R studying condition. The benefits of the R–T–R study condition were amplified at the 1-week delay assessment interval, paralleling well-established laboratory findings (e.g., Karpicke and Blunt 2011; Roediger and Karpicke 2006) and exceeded the benefits of the R–R + N strategy. With regard to the second objective, students who were shown the results of the first experiment, illustrating the benefits of a testing strategy, and who were urged to use the strategy when preparing independently for subsequent course exams, showed superior performance compared to a different group of students taking the same exams without explicit knowledge of the benefits of a testing strategy.

The results of this two-phase investigation show that the testing effect is applicable to university-level science classrooms and, perhaps most importantly, that students may be able to overcome pre-established notions regarding what memory strategies are most effective for learning. In other words, the university-level students in this experiment showed signs that they are able to overcome their own inaccurate understanding of how best to learn, particularly when materials need to be remembered in the long term. Metacognitive research has shown that many university students' perceptions of ideal study strategies, and perhaps more to the point, the nature of long-term memory, are flawed (e.g., Hartwig and Dunlosky 2012; Kornell and Bjork 2007; Roediger and Karpicke 2006). For example, many students have the sense that the easier processing is during the learning phase, the easier it will be to retrieve the information at exam time (e.g., Alter and Oppenheimer 2009; McCabe 2011). The current study results are encouraging and extend this literature because they show that research participants responded to the evidence from Phase I of the study that the testing strategy facilitated superior learning performance and then potentially applied the testing strategy to future test preparation in the same course. University-level science instructors are encouraged to demonstrate the benefits of testing, perhaps by conducting experiments in their own classes (e.g., Einstein et al. 2012), to increase students' metacognitive awareness of appropriate learning strategies. Students at

this level may be convinced that strategies counterintuitive to their beliefs actually work if they are shown concrete data that contradicts those beliefs.

With respect to this study's findings compared to other laboratory and applied studies of the testing effect, the results are mostly consistent. The R–T–R strategy in this study was superior for facilitating immediate and delayed recall after one week compared to the R–R–R strategy, and the R–T–R emerged as superior to the R–N + R strategy at the delayed assessment point. The atypical finding that resulted from this investigation is that the R–R–R strategy did not benefit students at the immediate assessment point, but the R–T–R strategy did. Typically, less cognitively demanding strategies such as simple repeated review of materials, such as in the "cramming for an exam" example, show immediate benefits in terms of recall performance (e.g., Hartwig and Dunlosky 2012; Rawson and Kintsh 2005). In contrast, strategies that are more cognitively demanding, and create desirable difficulties, do not always show immediate benefits (Bjork 1994). Desirable difficulties make initial learning more difficult but benefit the learner in terms of long-term recall as shown by many testing effect studies (e.g., Bjork 1994; Karpicke and Blunt 2011; Roediger and Karpicke 2006). The unique finding that the R–R–R condition did not yield immediate recall performance that was superior to the R–T–R condition could be evidence that the complexity of the physiology and anatomy materials necessarily called for more active studying than the R–R–R strategy facilitates. Expository text materials are usually new to students so they do not have familiarity in the form of prior knowledge to help them learn the materials upon initial exposure and they are more complex because they contain multiple ideas and atypical sentence structures. A study that was performed on a variety of age groups and also used expository text passages found a similar advantage for the testing effect condition at both a 5-min and a 2-day delay (Meyer and Logan 2013). So it could be that when participants learn materials that are less complex (e.g., word lists), then the testing effect is not evident at an immediate assessment point. In contrast, it could be that when participants are learning more complex materials (e.g., science or expository texts) that the testing effect is indeed present at both immediate and delayed assessment points. All in all, the findings of the current study replicate and extend previous work on the testing effect. It confirms that in the long run, which is arguably the most important in terms of building on one's knowledge permanently, testing is the superior strategy and also shows that students are capable of overcoming metacognitive difficulties to employ the strategy independently.

What do these results mean for the theory of desirable difficulties, that is, that encoding strategies that are cognitively demanding may negatively affect short-term but enhance long-term retention (e.g., Bjork 1994; Bjork and Bjork 2011)? The current results certainly support the tenet that desirable difficulties benefit long-term recall as the R–T–R strategy was superior to both the strategy that involved passive review (R–R–R) and fairly automated/passive note taking (R–R + N). Future investigations may need to examine further whether complex materials, such as expository text materials, can ever be learned well with passive strategies in the short term. So, the current results may challenge the part of the theory that suggests desirable difficulties only enhance long-term recall.

Alternatively, a partial lack of theoretical support for desirable difficulties, based on the results of the current study, could be attributed to a limitation in the study's design or procedures. Specifically, the time students were allowed to implement the R–R–R condition was not controlled across research participants. So it could be some research participants capitalized on the lack of time constraints and were able to use more sophisticated learning strategies than simple reviewing or rereading. In other words, unequal time constraints could have allowed active strategies to be implemented by some participants.

Future studies will seek to replicate this finding with similar materials and more strict time limits to study materials. Additionally, future studies could attempt to take this literature further by examining the benefits of the testing effect on other learning skills, not just recall. For example, future studies could replicate the impact of the testing effect on recall, as in this study, and seek to extend findings to more complex learning processes such as comprehension, reasoning about learned materials, and the synthesis of ideas.

To conclude, the current study illustrates that, when cognitive science principles are applied to a university-level science course, such as anatomy and physiology, learning may be enhanced. If more science instructors were aware of cognitive science findings and understood themselves the psychological nature of learning and memory, they could more easily implement techniques in their courses and pass along the techniques to science students. Perhaps in training science instructors, more intensive course work in the area of cognitive science should be required. This suggestion may serve as an important avenue for enhancing the learning and success of science majors so that they are well prepared for careers in allied health and medical fields.

Acknowledgments None.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating fundamental concept and changing study habits. *Teaching of Psychology, 39*, 190–193.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772–775.

Karpicke, J. D., & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science, 15*, 966–968.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 6*, 19–24.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449–468.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Medical Education, 42*, 959–966.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. *Medical Education, 43*, 1174–1181.

Lipowski, S. L., Pyc, M. A., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. *Developmental Psychology, 50*, 994–1000.

Logan, J. M., Thompson, A. J., & Marshak, D. W. (2011). Testing to enhance retention in human anatomy. *Anatomical Sciences Education, 4*, 243–248.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*, 94–97.

McCabe, J. A. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 29*, 222–233.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399–414.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*, 516–522.

Meyer, D. H., & Logan, J. M. (2013). Taking the testing effect beyond college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*, 142–147.

National Academy of Sciences. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future. Report from the Committee on Prospering in the Global Economy of the 21st Century*. Washington, DC: The National Academies Press.

National Commission on Mathematics and Science Teaching for the 21st Century. (2000). *Before it's too late*. Washington, DC: U.S. Department of Education.

Rawson, K. A., & Kintsh, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*, 70–80.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*, 382–395.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*, 429–442.

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia, 51*, 2360–2370.