RICE UNIVERSITY

Do older adults benefit from effortful retrieval?

by

Alda G Rivas

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Masters of Arts

APPROVED, THESIS COMMITTEE:

Jessica M Logan, Chair
Assistant Professor of Psychology

Margaret E Beier
Associate Professor of Psychology

Simon Fischer-Baum
Assistant Professor of Psychology

HOUSTON, TEXAS
MAY 2015
Do Older Adults Benefit From Effortful Retrieval?

by

Alda G Rivas

The testing effect is the finding that memory benefits are higher after taking a test than after re-studying. This effect has been attributed to a higher level of effort to retrieve information during a learning phase. In the present study, younger and older adults from the community read four essays and then either re-read the essay or took a test. The tests differed in the level of effort required to recall the information. Two days later, participants took final short answer tests for all essays. The percentage of correct recall was higher for the low-effort conditions (initial multiple-choice) compared to the re-study condition (testing effect). No testing effect was found for the short answer condition. These results indicate that, in a sample of participants from the community, increasing retrieval effort does not always produce greater enhancements to learning. Multiple-choice tests can still be highly effective and convenient boosters for learning.
Acknowledgements

During my years at Rice, different people provided me with the academic and/or moral support that allowed me to complete this project. In the academic realm, I would like to thank my current advisor Jessica Logan for all the guidance and support these past years, but above all, for giving me the opportunity to start this project in the first place. I would also like to thank the rest of my committee: Dr. Margaret Beier and Dr. Simon Fischer-Baum for their helpful insight and feedback. Special thanks to my undergraduate advisor Cody Cox for all the time, guidance and cheers before and during this journey.

I would also like to thank the people at the Department of Psychology at Rice University. To the faculty, especially to Dr. David Lane for the all the time and knowledge shared with me. To Lanita and Jennifer for showing me all the ropes a person should know during graduate school, and to my classmates Carmen, Curtiss, Hao, Heather, and Kim, for their time, knowledge, and friendship.

Warmest thanks to my friends outside the department: Ahmad, Gaby, Hussain, Jojo, Neto, Shawna, Thuy, Wael, Yessica, and especially Karina and Milivoje for all their love and support. Last, but definitely not least, thanks to my parents Carlos and Alda, and my brothers Carlitos, Oto, and Vale for all their love, patience, support and understanding.
# Table of Contents

Acknowledgements............................................................................................................ iii

Table of Contents................................................................................................................ iv

List of Tables....................................................................................................................... v

List of Figures..................................................................................................................... vi

List of Appendices............................................................................................................... vii

Chapter 1: Introduction...................................................................................................... 1

  Testing Effect................................................................................................................... 1

  Why Might Testing Be Better Than Re-Studying?............................................................ 1

  Practical Application of the Testing Effect.................................................................... 5

  Older Adults and Continuing Education....................................................................... 6

  Episodic Memory in Older Adults................................................................................. 7

  Testing Effect in Older Adults....................................................................................... 8

  Purpose of the Present Study......................................................................................... 10

Chapter 2: Method............................................................................................................ 13

  Participants..................................................................................................................... 13

  Materials......................................................................................................................... 13

  Procedure......................................................................................................................... 15

Chapter 3: Results............................................................................................................. 17

Chapter 4: Discussion...................................................................................................... 39

  Conclusion......................................................................................................................... 44

References......................................................................................................................... 45
List of Tables

Table 1. Descriptive Statistics for Younger Group by Education Level…………….18
Table 2. Descriptive Statistics for Older Group by Education Level…………………19
Table 3. Correlations between age, education, and previous knowledge……………20
Table 4. Descriptive Statistics for Forgetting Rates By Condition, Group, and Education
   Level……………………………………………………………………………………………………24
Table 5. Correlations between initial and final self-efficacy and performance on final
   tests……………………………………………………………………………………………………27
Table 6. Descriptive Statistics for Initial Self-Efficacy by Age, Group, and Education
   Level…………………………………………………………………………………………………28
Table 7. Descriptive Statistics for Self-Efficacy After Final Test By Age, Group, and
   Education Level……………………………………………………………………………………28
Table 8. Descriptive Statistics for Self-Efficacy by Testing Format, Group, and
   Education Level……………………………………………………………………………………37
List of Figures

Figure 1. Mean correct recall for all conditions by phase………………………………..22
Figure 2. Average forgetting rates (proportion) for the different testing conditions……..25
Figure 3. Scatterplot of the relationship between initial self-efficacy and performance on
the final test for the multiple-choice condition………………………………….29
Figure 4. Scatterplot of the relationship between initial self-efficacy and performance on
the final test for the MCG condition………………………………….30
Figure 5. Scatterplot of the relationship between initial self-efficacy and performance on
the final test for the short answer condition………………………………….31
Figure 6. Scatterplot of the relationship between initial self-efficacy and performance on
the final test for the re-study condition………………………………….32
Figure 7. Scatterplot of the relationship between self-efficacy after completing the study
and performance on the final test for the multiple-choice condition………………33
Figure 8. Scatterplot of the relationship between self-efficacy after completing the study
and performance on the final test for the MCG condition………………….34
Figure 9. Scatterplot of the relationship between self-efficacy after completing the study
and performance on the final test for the short answer condition………………35
Figure 10. Scatterplot of the relationship between self-efficacy after completing the study
and performance on the final test for the re-study condition………………….36
Figure 11. Mean scores for self-efficacy after each of the conditions during the initial
session……………………………………………………………………..38
List of Appendices

Appendix A. Reading Materials.................................................................49
Appendix B. Testing Materials...............................................................53
Appendix C. Self-Efficacy for Memory..................................................55
Appendix D. Demographic Information Questionnaire..........................56
Do older adults benefit from effortful retrieval?

Testing Effect.

The testing effect refers to the finding that taking a test produces significantly higher memory accuracy when compared to memory for information that was just re-studied (e.g., Carrier & Pashler, 1992; Phelps, 2012). The general paradigm used in studies exploring the testing effect involves four parts: 1) the presentation of the information to be learned, 2) an initial test and a control condition (re-study), 3) a retention interval (which can vary from minutes to days), and 4) a final recall test. Results using this paradigm generally show significantly higher retention of material for information that was initially tested than for information that was initially re-studied. This effect has been found in the literature as early as 1910, but it only appeared sporadically in the literature between 1917 and 1992 (Karpicke & Roediger, 2008). Conversely, the last two decades have seen an increased interest in the testing effect, with researchers exploring its practical relevance for academic settings and its theoretical implications for models of memory.

Why Might Testing be Better Than Re-studying?

One prevailing theoretical explanation for the testing effect was the transfer-appropriate hypothesis (Morris, Bransford, & Franks, 1977). This hypothesis predicts that information has a higher probability of being recalled when the context of the final test matches the context of the practice test. Empirical support for this hypothesis was obtained by McDaniel and Fisher (1991) who found that participants showed higher accurate recall when questions in initial and final tests shared more features. Furthermore, McDaniel, Kowitz, and Dunay (1989) also provided support for a transfer-appropriate
explanation in a study that manipulated the cues given to participants during the initial and final cued-recall tests. The conditions were either compatible (phonemic-phonemic or semantic-semantic) or incompatible (phonemic-semantic, semantic-phonemic). As a control condition, the researchers included a group that did not receive an initial test. Their results indicated that the compatible condition produced higher recall than both the incompatible and the control conditions, with no significant difference in memory performance for the latter two conditions.

More recent research has considered the concept of elaborative processing as an explanation for the testing effect. Elaborative processing requires a higher cognitive effort during the process of initial encoding or initial retrieval of the to-be learned information. According to Karpicke and Blunt (2011) and Karpicke and Smith (2012), elaborative encoding is not the mechanism responsible for the memory advantages produced by testing. While the authors acknowledge the memory benefits of elaborative encoding, they also found that after a successful recall, elaborative encoding does not add memory advantages after the information has been successfully recalled (Karpicke & Smith, 2012). This finding makes elaborative encoding an unlikely explanation for the memory benefits produced by recalling the information. However, the exact mechanism through which testing or retrieval practice produces memory benefits is still an open question.

Another alternative is the effortful retrieval hypothesis. Support for the effortful retrieval explanation was provided by Carpenter and DeLosh (2006). In their study, the researchers tested predictions made by the transfer appropriate hypothesis and the effortful retrieval explanation. Effortful retrieval is the idea that fewer cues provided to
retrieve information during the initial session will result in higher correct recall after a
delay (Carpenter & DeLosh, 2006). The researchers required participants to learn lists of
words. Then some participants re-studied the words or completed a test. The formats for
the initial tests were recognition, cued recall, and free recall. This initial test was
manipulated within subjects, but for the final test, the testing format was between
subjects. This means that during the final memory test, some participants completed a
recognition test, other participants completed a cued recall test, and a third group
completed a free recall test. According to the transfer-appropriate hypothesis, final test
performance should have been higher for conditions in which both the initial and final
tests were the same format. According to the effortful retrieval explanation, the fewer
cues encountered during the initial phase, the more accurate recall in the final test. The
results supported the effortful retrieval hypothesis: greater accuracy was observed for
conditions in which free recall (fewer cues) was involved in the initial phase, regardless
of the format of the final test (Experiment 1).

In order to ensure that these results were due to reduced number of cues, and not
because other qualitative differences between testing formats, Carpenter and DeLosh
(2006) conducted a second experiment in which participants were again required to
memorize lists of words. The researchers manipulated the amount of cues given to
participants during the initial tests (as measured by number of letters provided before
successful recall of the word). The authors found that words recalled with fewer cues
during the initial tests were more likely to be recalled during the final test.

Another study that explored the effortful retrieval hypothesis was Carpenter
(2009). In her study, participants were asked to rate the level of association between pairs
of words (cue-target) and then they were either allowed to re-study the pairs (by rating them again), or received an initial cued test. After a performing a distractor task, participants were asked to recall as many targets as possible. The level of association between cue-target pairs was either weak (basket-bread) or strong (toast-bread). Their results indicated that strong cues produced a better recall during the initial condition, but “the advantage for strong cues was eliminated or reversed at the time of the final test” (Carpenter, 2009, p. 1566). Carpenter (2009) interpreted these results as support for the effortful retrieval hypothesis and proposed that mechanism behind it can be interpreted through a spreading activation theory. This theory relies on the concept of interconnection of nodes in long-term memory. During the initial test, the lack of association between a pair of words (basket – bread) will prompt subjects to produce an association between them by linking them with propositions. During the final test, a search in memory for the correct answer will activate the propositions that connect the concepts, resulting in a wider semantic network with more routes to arrive at the correct answer.

The concept of effortful retrieval has also been proposed by Schmidt and Björk (1992) as explanation for the testing effect, which is considered one of the several procedures that provide “desirable difficulties.” According to Schmidt and Björk (1992), strategies that promote higher performance during learning do not necessarily translate into high performance at a later time. On the other hand, desirable difficulties, which involve training strategies that are more difficult and do not produce the highest possible performance during training (spaced studying, testing, etc.) will produce higher performance after a delay. In other words, making learning challenging has bigger long-
term benefits than strategies that make learning easier. In that case, it might be that testing is more effective than re-studying because it increases effort during learning.

In agreement with Carpenter and DeLosh (2006) and Schmidt and Björk (1992), in the present study, I propose that a low number of cues during a practice test constitutes a desirable difficulty because working with fewer cues encourages learners to engage in effortful retrieval. Testing formats that provide a higher number of cues (e.g., multiple-choice) should result in higher performance during training, and lower performance after a delay. On the other hand, testing formats that provide a low number of cues (e.g., short answer) should produce lower performance during training but higher performance after a delay.

**Practical Application of the Testing Effect**

Our dependence on long term memory in our daily activities makes the testing effect a practical tool for a wide range of situations which require us to retrieve information. One obvious situation where the testing effect can be advantageous, and one which researchers have frequently targeted, is the classroom setting. However, instructors employ testing to varying degrees, from daily pop quizzes to a single final exam. Many use it solely for assessment and evaluation purposes (e.g., assigning grades) rather than using it as a powerful learning tool.

The reasons for this lack of incorporation may be that the stimuli used in most research involving the testing effect do not involve educational-relevant material (Roediger & Karpicke, 2006). The reasons may also be the same as those addressed by Dempster (1988) regarding the spacing effect (little is known about the effect to justify its widespread application, the effect is not sufficiently understood, etc.) or it may just be a
tradeoff between memory benefits and practical benefits when choosing which type of test to use. For example, multiple-choice tests are a popular choice with instructors, especially in large classes, because they are fast for students to take and easy for instructors to grade.

However, in line with elaborative processing theories, multiple-choice tests would produce limited memory benefits: recognizing the answer among the lures may not motivate participants to engage in elaborative retrieval. On the other hand, short answer tests take longer to answer and longer to grade but may encourage more elaboration of the semantic network in order to generate the appropriate response. A possible solution to this tradeoff may be to produce a format of testing which incorporates the benefits of each test format. This could potentially make the testing effect a more attractive tool for instructors, and justify the incorporation of the testing effect into the regular educational curriculum.

**Older Adults and Continuing Education**

Most studies of the testing effect have concentrated on typical college age students (e.g., 18-22) or younger students in high school or middle school. In the present study, we focus on adults who have an interest in furthering their education but who are older than the traditional age for college. Continuing education programs are directed towards adults who are considered non-traditional students. Students may be considered “non-traditional” in many ways, including not enrolling in undergraduate education immediately after earning a high school diploma, being financially independent from parents, and working full time during the academic year (U.S. Department of Education, 2002). The courses offered in continuing education programs are varied and range from
personal development (e.g., playing the piano) to obtaining certification in order to satisfy the minimum requirements for a specific job position (e.g., paralegal).

Depending on the age of the student, the reasons for returning to the classroom may vary. A younger adult may be interested in improving financial security; an older adult may be interested in filling free time after retirement. Additionally, with longer life expectancies, an older adult may decide to continue his or her education in order to postpone retirement due to financial concerns.

**Episodic Memory in Older Adults**

One potential issue facing older adults in the classroom is that their cognitive abilities are often different than those of their younger counterparts. While older adults’ performance in some cognitive abilities may remain constant (procedural memory), or even improve over time (semantic memory), other cognitive abilities decline with age (episodic memory). Procedural memory, which involves a wide range of knowledge (including skills), has the characteristic of not requiring a conscious effort to access the knowledge (e.g., how to write). Reports in the literature indicate that this type of memory remains relatively intact as age progresses (e.g., Balota, Dolan, & Duchek, 2000).

Semantic memory, which involves general knowledge and facts about the world (e.g., the name of objects) not only remains constant, but it has also been found to increase with age (Parks, Polk, Mikels, Taylor, & Marshuetz, 2001).

Episodic memory, on the other hand, has consistently been reported in the literature as declining with increased age (Balota, Dolan, & Duchek, 2000). One characteristic of information that is considered episodic in nature involves a conscious effort to recall such information, such as a test that will require recall of information
learned in lecture. Laboratory studies typically show that older adults will score lower than younger adults on recall test (Balota, et al., 2000), particularly during intentional learning (Naveh-Benjamin, Shing, Kilb, Werkle-Bergner, Lindenberger, & Li, 2009).

The good news is that the difficulty in recalling episodic information does not necessarily mean that such information is inaccessible. Converging evidence in the literature seems to suggest that when older adults are provided with cues about the information to be recalled, or when the context of the information to be remembered resembles the natural context where such information is found, the differences in performance between younger and older adults start to dissipate (Castel, 2005). For example, in a study that required participants to remember prices of groceries, older participants performed worse than younger adults when the prices of the items were unusual prices compared to what shoppers would normally find at the grocery stores (Castel, 2005). However, when the prices of groceries resembled the prices in grocery stores (e.g., they had context that made sense), older adults performed as well as younger adults. Similarly, in testing, when the format of a test provides cues about the correct answer (e.g., multiple-choice or cued recall) older participants perform better than when the test does not provide such cues (e.g., free recall) (Balota, Dolan, & Duchek, 2000).

**Testing effect in older adults**

Since more non-traditional students are enrolling in continuing education classes every year (U.S. Department of Education, 2012), it seems worthwhile to explore whether the benefits of testing would generalize to this non-traditional population. A handful of studies in the literature have already explored this idea. However, issues with
the study design (lack of re-study control) prevent a strong conclusion in favor of testing over study (Henkel, 2007, 2008; Logan & Balota, 2008).

The issue of lacking a re-study control group in a study exploring the effects of testing in older adults was addressed by Tse, Balota, and Roediger (2010). In their study, the authors explored the repeated study and repeated testing effects in memory recall of face-name pairs in middle age and older adults. The study was divided in two experiments, only one of which included feedback. Their findings indicated that in the no-feedback condition, middle age adults (Mean age = 61 years) benefited more from repeated testing, whereas older adults (Mean age = 72 years) benefited more from repeated study. However, when feedback was provided (Experiment 2), a testing effect was also found in the older adult sample (Mean age = 80 years). While this study provides strong evidence in favor of the benefits of testing in older adults, the materials used as stimuli (face-name pairs) do not allow for a strong generalization to incorporate repeated testing as a learning tool in continuing education classrooms.

Meyer and Logan (2013) explored the testing effect in older adults using more educationally relevant materials. In this study, the materials used were one-page encyclopedia articles about varied subjects (human heart, tsunamis, black holes, and armadillos), which made the material to be remembered more similar to the prose encountered in textbooks. The design of the study included a re-study condition that served as control, and it also allowed as much re-exposure time to the information to be learned as the testing condition did. Furthermore, the authors explored two age groups (18-25 and 55-65). The younger group included 2 subgroups: one of them composed of university students, and the other one composed of young adults from the community.
This made their sample more representative of the age range of the population enrolled in an academic setting (traditional college or continuing education). Not surprisingly, the sample of university students outperformed both community groups. Importantly, however, older participants benefited from the testing conditions as much as the younger participants from the community did.

**Purpose of the Present Study**

The first purpose of this study was to find a test format that maximizes the memory benefits provided by repeated testing, and is practical enough in a classroom context. A second purpose was to explore whether the memory benefits provided by different test formats are the same for younger and older adults. In order to achieve this, I attempted to replicate and extend the findings from Meyer and Logan (2013) and Tse, Balota, and Roediger (2010), but with a sample that may be more representative of an age range of older adults interested in academic pursuits (at least 50 years old and up to 70 years old). More specifically, I explored whether older adults can benefit from repeated testing when compared to restudying, and whether the level of effort required to retrieve the answer during an initial test has an effect on later recall. Following the results of literature reviewed above, I predicted that:

**Hypothesis 1**: Both older and younger adults would show memory benefits from testing conditions in comparison to the re-study condition (both groups would show a testing effect).

I also explored whether both older and younger adults would benefit to the same extent from the different levels of desirable difficulties afforded by different testing formats. According to the effortful retrieval hypothesis, fewer cues provided by a
question in order to retrieve the correct answer will result in a more elaborate network of potential retrieval routes compared to questions that provide more cues. Therefore, I expected that:

*Hypothesis 2.* Completing a short answer test during the initial session would result in more memory benefits during the final test compared to having a multiple-choice test during the initial session for both older and younger adults.

In order to explore whether participants can benefit from combining an effortful level of retrieval (short answer question) followed by the provision of cues (multiple-choice question), I included a format of testing that I call “Multiple-choice generate” (MCG). This MCG format will be described in more detail in the method section of this paper. Because MCG will provide both low and high cues during the initial test, I make no specific predictions about the final performance for this condition, and the analysis performed on this data is exploratory.

Another purpose of this study was to explore participants’ ratings of self-efficacy to remember information. According to Bandura (1977), one of the sources of self-efficacy is the experience that the individual has had in a specific situation. In lack of previous experience (e.g., a novel situation), self-efficacy is determined by generalizing from previous situations that share characteristics with the novel situation. However, as the person engages in the novel situation and observer his/her own performance, self-efficacy will be adjusted to reflect either the perceived success or failure in the new situation. In the context of the present study, I expected participants’ initial self-efficacy to reflect the self-efficacy generalized from prior experience in activities involving
memory and learning. However, as participants advance through the task and acquire new information and experience about their performance, self-efficacy will change based on whether they perceive that their performance is being successful or not.

*Hypothesis 3a.* Participants’ final self-efficacy will be more highly correlated with performance than participants’ initial self-efficacy.

As participants complete the tests during the initial session, they might realize that they cannot remember all the answers to the questions. On the other hand, after re-studying the information during the initial session, participants will not take a test and therefore, will not be able to estimate how much information they have forgotten. This may lead participants to overestimate the amount of information that they will remember after re-studying and to underestimate the amount of information that they will remember after a test, and this will be reflected in participants’ reports of self-efficacy.

Previous studies have also have compared participants’ judgments of learning after testing and after re-studying. Judgment of learning is a concept similar to self-efficacy. It involves participants reporting the percentage of information that they believe will be able remember in a final test. According to Karpicke and Blunt (2011), people usually associate re-studying with learning, and testing with measuring learning. In their study, they found that participants consistently expect to remember more information after re-studying material than after being tested. In light of these previous findings, I expected that participants’ memory self-efficacy at the end of the initial session would be rated higher for re-studying than for taking a test. Among the three testing conditions, I anticipated participants’ ratings of self-efficacy to be in line with my predictions for their performance: conditions that provide more cues during the learning phase (e.g., multiple-
choice) would result in higher performance, and therefore, in higher self-efficacy. On the other hand, conditions that provide fewer cues (e.g., short answer) would result in lower initial performance, and this experience will result in lower self-efficacy.

*Hypothesis 3b.* Participants would report higher rating of self-efficacy after re-studying compared to the self-efficacy ratings for the testing conditions and would report lower ratings of self-efficacy after completing a short answer initial test than after completing a multiple-choice initial test.

Because the MC-G test involves both multiple choice and short answer questions, I make no predictions regarding the participants’ self-efficacy for this condition.

**Method**

**Participants**

Sample consisted of 41 older adults (ages 50-70) and 40 younger adults (ages 18-30), who were recruited through fliers posted in the community. In addition to age, the requirement for participation was to be a fluent English speaker and to have basic computer experience. Even though the sample of older adults from Meyer and Logan (2013) consisted of adults 55 and above, our reasoning for recruiting somewhat younger participants is that 50+ years of age has been referred to in the literature as the beginning of a noticeable decline in memory performance (Erngrund, Mäntylä, & Nilsson, 1996). The requirement of being a fluent English speaker was to ensure that participants were able to understand the text presented to them during the task.

**Materials**

The materials included four one-page articles (324 to 577 words) adapted from National Geographic Online [as used by Meyer and Logan (2013); see Appendix A].
Each article had a different subject: armadillos, black holes, human heart, and tsunamis. For each article, three 10-question tests were created for use in the initial phase. Participants only received one test for each article and test type was counterbalanced across participants. Each of the three tests had the same 10 questions presented in different format, depending on the testing condition: multiple-choice typical, multiple-choice generate, and short answer. The stem of the question was the same for all the conditions, but they differed individuals responded. For example, after reading the armadillo passage, the test in the short answer condition included the question “Which type of armadillo can contort itself into a hard ball to thwart predators?” The question was followed by a blank area for participants to enter their response. For the multiple-choice typical condition, participants saw exactly the same question, but they were presented with four choices (one correct answer and three lures). For the multiple-choice generate (MCG) condition, participants saw the short answer version of the question, and only after generating a response, participants were presented with the same four choices as in the multiple-choice condition, from which they selected the answer that they considered to be the correct one.

The articles were counterbalanced across the three testing conditions and one re-study condition (control) for a total of four counterbalances. For the final phase, there were four tests in total. These tests included the same questions that were seen during the initial tests, plus a test for the re-study condition. All the final tests were presented in short-answer format (tests can be found in Appendix B).

Three adaptations from Quiñones (1995) self-efficacy scale were included int his study (Appendix C). One scale included questions about memory in general and its
purpose was to measure participants’ self-rated ability to learn the information they were about to read. This scale consisted of ten items, an example of which is “I feel confident in my ability to learn the information I am about to read.” Another ten-item scale was included in order to assess self-efficacy for memory after training. This scale was practically identical to the scale described above, with the exception that it aimed at measuring participants’ self-rated ability to retain the information learned after training. An example of an item from this scale is “I feel confident in my ability to remember this information in the future.” Finally, a 5-item self-efficacy scale was incorporated to the study in order to measure self-rated memory for each article. An example of an item from this scale is “I feel confident in my ability to remember the information from this article during the final test.” Participants responded to all the self-efficacy questions by selecting an option from a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). In addition to these materials, participants also completed a questionnaire that included demographic questions such as age and highest level of education achieved, as well as questions about their previous knowledge for each of the articles (Appendix D).

**Procedure**

After listening to a general description of the experiment, participants read and signed the informed consent, and were then asked to sit in front of a computer. In order to ensure that participants would not feel pressured to perform faster than their own speed, all participants completed the study in a room with only one computer. Before starting the experiment, participants received instructions to pay attention and only read each article once, at their own pace.
After reading the first article, participants completed multiplication problems for 5 minutes and then either answered a test (multiple-choice typical, short answer, or multiple-choice generate format) or re-read the article, depending on the counterbalance to which they were assigned. Each question in the tests was presented by itself on the screen. Participants were not able to navigate back through the experiment, and were not able to go forward, until an answer had been entered for the question currently on the screen. After answering the last question in a test, participants completed the self-efficacy scale for the article they had just read before the test, and then they saw the next article. This process was repeated three times (until participants had read the four articles, answered the test or re-read the information, and answered the self-efficacy scale for each article). This completed the initial session on the study, and participants were then asked to provide the experimenter with an email address. The purpose of asking for this information was to send the participants a reminder email one day before the second session. Participants were instructed to not try to rehearse or google the information seen in the articles.

Two days later, participants returned for the final session to complete the final recall test. This test consisted of 40 short answer questions (the same 30 questions from the tests in learning phase, plus 10 questions covering the re-study section). After completing the final test, participants completed the last self-efficacy scale, which measured participants’ judgment of their ability to remember the information at a later time. Participants were then compensated $50 for their participation.
Results

The dependent measure for this study was the percentage of correct responses for each test. Means and standard deviations by test, phase, and education level can be found in Table 1 (younger adults) and Table 2 (older adults). Correlations between age, education level, and previous knowledge are presented in Table 3. Two different raters independently scored the responses for all the short answer questions. The inter-rater reliability for this data was assessed by calculating an inter-class correlation coefficient using a two-way random model. The results of this analysis indicated that the absolute agreement between the two raters was excellent ($ICC = .97, p < .001$).
Table 1

Descriptive statistics for younger group by education level

<table>
<thead>
<tr>
<th></th>
<th>High school or less</th>
<th>College degree</th>
<th>Graduate work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Multiple-choice (I)</td>
<td>85.4</td>
<td>13.3</td>
<td>78.7</td>
</tr>
<tr>
<td>Multiple-choice (F)</td>
<td>71.5</td>
<td>20.3</td>
<td>69</td>
</tr>
<tr>
<td>MCG (I)</td>
<td>68</td>
<td>23.8</td>
<td>74.3</td>
</tr>
<tr>
<td>MCG (F)</td>
<td>65</td>
<td>18.1</td>
<td>62.3</td>
</tr>
<tr>
<td>Short answer (I)</td>
<td>56</td>
<td>20.8</td>
<td>54.33</td>
</tr>
<tr>
<td>Short answer (F)</td>
<td>50.5</td>
<td>26.4</td>
<td>49</td>
</tr>
<tr>
<td>Re-study (F)</td>
<td>53</td>
<td>19.7</td>
<td>49.7</td>
</tr>
</tbody>
</table>

Note. I = initial session. F = Final session. Total sample size for the younger group n = 40. High school or less n = 10, college degree n = 15, graduate work n = 15.
Table 2

Descriptive statistics for older group by education level

<table>
<thead>
<tr>
<th></th>
<th>High school or less</th>
<th>College degree</th>
<th>Graduate work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Multiple-choice (I)</td>
<td>69.6</td>
<td>19.9</td>
<td>69.8</td>
</tr>
<tr>
<td>Multiple-choice (F)</td>
<td>48.2</td>
<td>21.3</td>
<td>59.1</td>
</tr>
<tr>
<td>MCG (I)</td>
<td>63.8</td>
<td>27.8</td>
<td>71.1</td>
</tr>
<tr>
<td>MCG (F)</td>
<td>48.9</td>
<td>19.4</td>
<td>63.6</td>
</tr>
<tr>
<td>Short answer (I)</td>
<td>45.4</td>
<td>20.3</td>
<td>57.7</td>
</tr>
<tr>
<td>Short answer (F)</td>
<td>43.6</td>
<td>18</td>
<td>49.1</td>
</tr>
<tr>
<td>Re-study (F)</td>
<td>38.2</td>
<td>16.5</td>
<td>50</td>
</tr>
</tbody>
</table>

Note. I = initial session. F = Final session. Total sample size for the older group n = 41. High school or less n = 14, college degree n = 11, graduate work n = 16.
Table 3

*Correlations between age, education, and previous knowledge.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Education</td>
<td>-.04</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3. Previous Knowledge</td>
<td>-.12</td>
<td>.26*</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: N = 81. *p < .05.
Data were analyzed using a 2 (group: younger, older) x 2 (phase: initial, final) x 3 (testing condition: multiple-choice typical, multiple-choice generate, short answer) mixed design (age group as between subjects factor, phase and testing condition as within subjects factors) ANCOVA with previous knowledge and education level as covariates. This analysis revealed that previous knowledge had an effect that approached statistical significance ($p = .055$) and that education level did not have a statistically significant effect ($p = .127$). A statistically significant interaction was found between phase and testing condition: $F(2, 138) = 3.34, p = .038$, partial $\eta^2 = .046$ after controlling for previous knowledge and education level. The main effect of age group ($p = .64$) and test ($p = .09$) were not significant. The power to detect this interaction was .62. There were no other significant main effects or interactions. The interaction between phase and testing condition is represented in Figure 1. Planned contrasts were performed on the data in order to analyze the simple effect of testing across the two levels of phase.
For the initial phase, the contrasts indicated that participants scored lower on the short answer test compared to the multiple-choice and MCG tests ($t(80) = 8.47, p < .001, d = 0.94$), with no significant difference in performance between the multiple-choice and MCG conditions ($p > .05$). For the final phase, participants scored significantly lower in the re-study condition compared to the multiple-choice condition ($t(80) = 7.29, p < .001, d = 0.81$) and the MCG condition ($t(80) = 6.34, p < .001, d = 0.71$), but not significantly different than the short answer condition ($p > .05$). This means that there was a testing effect for multiple-choice and MCG conditions, but not for the short answer condition. Because not all testing conditions produced significantly higher scores than the re-study condition, these results offer partial support for Hypothesis 1.

Figure 1. Mean correct recall for all conditions by phase. Error bars indicate standard error of the mean.
In order to explore whether the having a more effortful test during the initial session produces bigger memory benefits (higher score in the final exam), planned contrasts were performed between the final scores for the testing conditions. The results indicated that completing an initial short answer tests results in significantly lower final scores than having an initial multiple choice \( t(80) = 6.19, p < .001, d = 0.68 \) or an initial MCG test \( t(80) = 4.66, p < .001, d = 0.51 \) with no significant difference between having an initial multiple-choice or an initial MCG \( (p > .05) \). Because we assumed that the higher effortful retrieval was produced by the initial short answer test, these results do not support Hypothesis 2.

In order to explore whether a higher level of effort to retrieve the information during the initial session provides more memory benefits (less forgetting), we computed new variables to represent the proportion of information forgotten by participants. Proportion of information forgotten was used instead of raw forget scores, as recommended by Roediger and Karpicke (2006). Two data points in the short answer condition and one data point in the MCG condition were identified as outliers (outside 3IQR) and were replaced with the grand mean. Table 4 shows the means and standard deviations for forgetting rates.
Table 4

*Descriptive statistics for forgetting rates by condition, group, and education level.*

<table>
<thead>
<tr>
<th></th>
<th>Multiple-choice</th>
<th>MCG</th>
<th>Short answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Less than college degree (Y)</td>
<td>.18</td>
<td>.18</td>
<td>.08</td>
</tr>
<tr>
<td>Less than college degree (O)</td>
<td>.32</td>
<td>.24</td>
<td>.21</td>
</tr>
<tr>
<td>College degree (Y)</td>
<td>.16</td>
<td>.16</td>
<td>.19</td>
</tr>
<tr>
<td>College degree (O)</td>
<td>.20</td>
<td>.19</td>
<td>.13</td>
</tr>
<tr>
<td>Higher than college degree (Y)</td>
<td>.16</td>
<td>.15</td>
<td>.12</td>
</tr>
<tr>
<td>Higher than college degree (O)</td>
<td>.16</td>
<td>.15</td>
<td>.16</td>
</tr>
</tbody>
</table>

*Note.* Y = younger group. O = older group. Total sample size for the younger group n = 40 (high school or less n = 10, college degree n = 15, graduate work n = 15). Total sample size for the older group n = 41 (high school or less n = 14, college degree n = 11, graduate work n = 16.)
A 3 (multiple-choice, MCG, short answer) x 2 (younger, older) x 3 (less than college degree, college degree, higher than college degree) mixed design ANOVA with testing condition as within subjects factor, age as between subjects factor, and education level as between subjects factor was performed on the data. The results indicated a significant main effect of testing condition \( (F(2, 150) = 6.41, \eta^2 = .079, p = .002) \). Power to detect this effect was .90. No other effects or interactions were significant. Figure 2 represents the average forgetting rates by condition.

![Figure 2](image.png)

**Figure 2.** Average forgetting rates (proportion) for the different testing conditions. Error bars represent standard error of the mean.

Post-hoc contrasts indicated that participants forgot significantly less information in the short answer condition than in either the multiple-choice \( (t(80) = 3.74, p < .001, d = 0.16) \) or the MCG delay condition \( (t(80) = 2.21, p = .03, d = 0.24) \) with no significant difference in forgetting rates between the multiple-choice and short answer conditions \( (p < .05) \). Although the literature conceptualizes the memory benefits of testing as high
performance in a final test, here we propose that protection against forgetting can also be considered a memory benefit. Therefore, we consider the analysis on forgetting rates as providing support for Hypothesis 2.

In order to explore the extent to which self-efficacy before and after completing the study was related to performance, a Pearson’s correlation was performed between each of the testing conditions and self-efficacy before and after completing the study. The results of this analysis indicated that self-efficacy before completing the study was not significantly related to any of the final tests ($p > .05$) and self-efficacy after completing the study was significantly related to all the final tests, providing support for Hypothesis 3a. Correlations for initial and final self-efficacy and score in the final tests are presented in Table 5. Means and standard deviations for the average self-efficacy by group and education level can be found in Tables 6 and 7. Figures 3 – 6 represent the relationship between self-efficacy (before participants completed the study) and performance. Figures 7 – 10 represent the relationship between self-efficacy (after participants completed the study) and performance.
Table 5

Correlations between initial and final self-efficacy and performance on final tests.

<table>
<thead>
<tr>
<th></th>
<th>Initial self-efficacy</th>
<th>Final self-efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$p$</td>
</tr>
<tr>
<td>Multiple-choice</td>
<td>.09</td>
<td>.45</td>
</tr>
<tr>
<td>MCG</td>
<td>.14</td>
<td>.20</td>
</tr>
<tr>
<td>Short answer</td>
<td>-.08</td>
<td>.50</td>
</tr>
<tr>
<td>Re-study</td>
<td>.10</td>
<td>.38</td>
</tr>
</tbody>
</table>

*Note. $r$ = correlation value. $p$ = probability.*
Table 6

Descriptive statistics for initial self-efficacy by age group and education level.

<table>
<thead>
<tr>
<th></th>
<th>Younger group</th>
<th></th>
<th>Older group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>High school or less</td>
<td>32.9</td>
<td>5.7</td>
<td>29.86</td>
<td>5.7</td>
</tr>
<tr>
<td>College degree</td>
<td>32.8</td>
<td>5.1</td>
<td>33</td>
<td>4.4</td>
</tr>
<tr>
<td>Graduate work</td>
<td>31.93</td>
<td>5.1</td>
<td>32</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Note. Total sample size for the younger group n = 40. Total sample size for the older group n = 41. For younger group: High school or less n = 10, college degree n = 15, graduate work n = 15. For older group: High school or less n = 14, college degree n = 11, graduate work n = 16.

Table 7

Descriptive statistics for self-efficacy after final test by age group and education level.

<table>
<thead>
<tr>
<th></th>
<th>Younger group</th>
<th></th>
<th>Older group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>High school or less</td>
<td>31.1</td>
<td>5.7</td>
<td>29</td>
<td>4.7</td>
</tr>
<tr>
<td>College degree</td>
<td>32.53</td>
<td>6.4</td>
<td>32.9</td>
<td>7</td>
</tr>
<tr>
<td>Graduate work</td>
<td>32.33</td>
<td>6.2</td>
<td>32.06</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Note. Total sample size for the younger group n = 40. Total sample size for the older group n = 41. For younger group: High school or less n = 10, college degree n = 15, graduate work n = 15.
Figure 3. Scatterplot of the relationship between initial self-efficacy and performance on the final test for the multiple-choice condition.
Figure 4. Scatterplot of the relationship between initial self-efficacy and performance on the final test for the MCG condition.
Figure 5. Scatterplot of the relationship between initial self-efficacy and performance on the final test for the short answer condition.
Figure 6. Scatterplot of the relationship between initial self-efficacy and performance on the final test for the re-study condition.
Figure 7. Scatterplot of the relationship between self-efficacy after completing the study and performance on the final test for the multiple-choice condition.
Figure 8. Scatterplot of the relationship between self-efficacy after completing the study and performance on the final test for the MCG condition.
Figure 9. Scatterplot of the relationship between self-efficacy after completing the study and performance on the final test for the short answer condition.
In order to explore whether the testing format (or re-studying the article) produced changes in participants’ self-efficacy, a 2 (self-efficacy) x 2 (group) x 3 (education level) mixed-design ANOVA was performed with self-efficacy for each testing format as within-subjects factor, and age group and education level as between subjects factors. Means and standard deviations for self-efficacy for each testing condition, by age group, and education level can be found in Table 8. Self-efficacy by testing condition collapsed across age groups and education level can be found in Figure 11.
Table 8

Descriptive statistics for self-efficacy by testing format, group, and education level.

<table>
<thead>
<tr>
<th>Test format</th>
<th>Group</th>
<th>High school or less</th>
<th>College degree</th>
<th>Graduate work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Multiple-choice (Y)</td>
<td>18.3</td>
<td>4.03</td>
<td>17.53</td>
<td>3.81</td>
</tr>
<tr>
<td>Multiple-choice (O)</td>
<td>16.86</td>
<td>3.68</td>
<td>17.55</td>
<td>3.27</td>
</tr>
<tr>
<td>MCG (Y)</td>
<td>16.30</td>
<td>5.77</td>
<td>16.13</td>
<td>3.38</td>
</tr>
<tr>
<td>MCG (O)</td>
<td>14.43</td>
<td>4.07</td>
<td>16.64</td>
<td>3.30</td>
</tr>
<tr>
<td>Short answer (Y)</td>
<td>16.20</td>
<td>5.61</td>
<td>15.40</td>
<td>5.15</td>
</tr>
<tr>
<td>Short answer (O)</td>
<td>15.29</td>
<td>3.02</td>
<td>14.55</td>
<td>4.91</td>
</tr>
<tr>
<td>Re-study (Y)</td>
<td>18.50</td>
<td>2.59</td>
<td>18.53</td>
<td>3.4</td>
</tr>
<tr>
<td>Re-study (O)</td>
<td>16.29</td>
<td>3.58</td>
<td>16.82</td>
<td>3.43</td>
</tr>
</tbody>
</table>

Note. Y = younger group. O = older group. Total sample size for the younger group n = 40. Total sample size for the older group n = 41. For younger group: High school or less n = 10, college degree n = 15, graduate work n = 15. For older group: High school or less n = 14, college degree n = 11, graduate work n = 16.
Figure 11. Mean scores for self-efficacy after each of the conditions during initial session. Error bars represent standard error of the mean.

The results of the ANOVA indicated a significant effect of testing condition on self-efficacy ($F(3, 225) = 11.52, p < .001$, partial $\eta^2 = .133$, power: .99). No other effect was statistically significant, indicating that there were significant differences in the reported self-efficacy by age group or education level. Planned pairwise comparisons with Bonferroni correction indicated that people reported higher levels of self-efficacy after completing a multiple-choice test than after completing either the MCG test ($p < .001$) or the short answer test ($p = .001$), higher self-efficacy after re-reading the information than after completing either the MCG ($p = .004$) or the short answer test ($p < .001$). No significant differences were found between completing a multiple-choice test and re-studying the information ($p > .05$) or between completing the MCG and the short answer test ($p > .05$). These results provide partial support for Hypothesis 3b.
Discussion

The purpose of this study was to replicate and extend the findings of Logan and Meyer (2013). Specifically, we explored 1) whether a testing effect could be found with an older sample; 2) whether the benefits of testing compared to re-studying were the same for older than for younger adults (whether the magnitude of the testing effect was the same for both samples); 3) whether different levels of effortful retrieval during the practice test had an effect in later retention; 4) the relationship between initial and final self-efficacy with final performance, and 5) whether self-efficacy to remember the information in the final test would vary as a function of test format during the initial phase.

The results of the present study indicated that both younger and older adults do show a memory advantage for later recall after taking a test in comparison to re-reading the information. The fact that there was no significant main effect or interaction with group or education level indicated that both younger and older participants from different education backgrounds performed at the same level and benefited to the same extent from taking a test compared to re-studying. In other words, I replicated the results of Meyer and Logan (2013) who found a testing effect in a sample of older adults with no difference in performance in neither initial nor final phase when compared to a younger sample from the community.

Contrary to my predictions, however, neither of our groups showed the expected memory benefits of testing when the initial test was a short answer format. Final test performance after completing a short answer test was not different from the final test performance after re-studying. This may have been due to either the timing between
initial and final test, or the lack of feedback provided to participants during the initial test. It is possible that a longer retention interval might be necessary in order to observe the effect of testing after taking an initial short answer test. Perhaps with a longer retention interval, performance in a final test after re-studying the information might show a steeper decline than performance after taking a final test. The other potential explanation is the lack of feedback in this study after the initial test.

Tse, Balota, and Roediger (2010) explored the effect of feedback in later retention of face-name pairs in a sample of older adults. This sample was subdivided into three groups by mean age: 60.81 years, 71.97, and 81.28. In their first experiment, which provided no feedback, the authors found a testing effect for the group with a mean age of 60.81 years, but not for the other two groups. In their second study, which included feedback, the authors found a testing effect for all groups. Furthermore, it has been shown that feedback provides memory benefits above and beyond those provided by repeated testing (Roediger & Butler, 2010). However, for the purposes of the present study, we decided to not provide feedback in order explore the effects of different levels of effortful retrieval without combining them with the effect of feedback.

Regarding the role of effortful retrieval, our study extended the findings of Meyer and Logan (2013). While they presented participants with multiple-choice tests and re-study tasks during the initial phase, our study also included the MCG and short answer during the initial phase. The purpose of manipulating the format of testing during the initial phase was to explore whether a higher level of desirable difficulties (engaging in elaborative retrieval during the initial test) would result in higher performance during the final test. My predictions were that performance in a final test would be highest after
taking an initial short answer test, lower after a multiple choice test, and lowest after re-
studying.

In order to be able to test my predictions empirically, I worked under the assumption that the re-studying condition provides no opportunity to engage elaborative retrieval processes because no retrieval is required. In other words, I assumed that re-
studying provides no desirable difficulties. I also assumed that the short condition would provide the highest level of desirable difficulties because the only cue provided was the question itself. Therefore, this condition provides the highest opportunity to engage in elaborative retrieval processes. Finally, I assumed that the multiple-choice condition provided a higher level of desirable difficulties than re-studying, but not as high as short answer tests. This is because multiple-choice tests still require retrieval (contrary to re-
studying, which involves mainly input of information), but provides more cues to arrive to the answer than short answer tests.

In line with my predictions and with the elaborative retrieval hypothesis, taking a final test after just re-reading the information resulted in the lowest performance compared to taking the multiple-choice and MCG tests. However, contrary to my predictions, performance on a final test was not higher after taking an initial short answer test, and was actually as low as performance after re-studying. One potential explanation for this finding is, again, the fact that participants did not receive feedback. Perhaps recognizing an answer among lures provides participants with feedback, an advantage that is not available in the short answer test. This explanation seems feasible because performance in the final test following the MCG condition (which includes a short
answer question followed by the same question in multiple-choice format) was not different than performance in the multiple-choice test.

One more potential explanation for the lack of testing effect in the short answer condition may involve the level of effortful retrieval itself. If we organize the conditions in this study in a hierarchical order (from lowest to highest level of effort required for recall), then we have: re-study – multiple-choice – MCG – short answer. As stated in the result section, a testing effect was found for both the multiple-choice and MCG, but not for short answer. Notice that the conditions that resulted in testing effect are the ones that require a medium level of effort for recall. Perhaps there’s an optimal level of effortful retrieval in order to observe memory benefits compared to re-studying. If so, the memory benefits produced by effortful retrieval might increase with effort and would peak at some threshold, after which the memory benefits would decline.

Regarding the memory benefits of testing and desirable difficulties, my findings indicated that taking an initial short answer test led to lower forgetting rates than taking an initial multiple-choice or MCG test. This finding is line with previous findings in the literature regarding the advantages of introducing desirable difficulties. However, a potential explanation for this difference in forgetting rates might be the fact that performance on the initial short answer test was already lower than initial performance in multiple-choice and MCG tests, with the later two showing ceiling effects in the initial test.

Regarding differences in the reported self-efficacy after different initial conditions (multiple-choice, MCG, short answer, re-studying) my predictions were partially supported: I anticipated that participants would report higher self-efficacy for
conditions with lower levels of required effortful retrieval (re-studying < multiple choice < short answer). My results indicated, as predicted, that participants reported lower self-efficacy for short answer than for re-studying and multiple choice. Contrary to my predictions, participants did not rate their self-efficacy for multiple-choice as lower than re-studying.

These results partially resemble those of Karpicke and Blunt (2011). In their study, the researchers required participants to read a science text, and then manipulated the conditions for study. One of the conditions required participants to re-study the text, and another condition involved free recall. The researchers also collected ratings of self-efficacy by requiring participants to provide their judgment of learning. The results indicated that participants rated their learning as higher for the re-study condition than the free recall condition, even though their actual performance showed the opposite pattern.

Finally, the purpose of introducing the MCG was to include a test that combined the desirable difficulties during a learning phase (as recommended by Schmidt & Björk, 1992) with the practicality provided by multiple-choice tests (which are easier and faster to grade). The MCG test consisted of presenting questions in short answer format (low level of cues) first in order to encourage participants to engage in effortful retrieval. Only after an answer had been generated, participants were presented with the correct answer among three lures, and then required to select their response. The analysis indicated that participants showed similar magnitude of testing effects for the MCG and multiple-choice condition, as indicated by their effect sizes ($d = 0.81$ and $0.71$, respectively). Regarding the ratings of self-efficacy, however, the results indicated that participants’ self-efficacy to remember the information was as low as the ratings of self-efficacy for
the short answer test. This is in line with previous findings, which suggest that participants do not have accurate judgments of their future performance in different conditions (Karpicke & Blunt, 2011; Tse, Balota and Roediger III, 2010).

**Conclusion**

The findings of the present study indicate that multiple-choice tests promote greater memory benefits for younger and older samples than short answer tests and re-studying. Furthermore, multiple-choice tests result in higher self-efficacy than MCG and short answer tests. This may have important implications in students’ decision to keep pursuing educational goals. Future studies may benefit from exploring the effect of feedback on learning and self-efficacy across different levels of effortful retrieval. Self-efficacy ratings for re-study should resemble more the actual performance after receiving feedback. This is an important issue because it provides the potential to change students’ erroneous idea that re-studying information would provide them with more memory benefits than taking a test. Allowing students to compare their actual performance after a test and after re-studying may motivate them to stop seeing a test as only a tool to measure knowledge and start seeing it as tool for learning.

Exploring longer retention intervals between initial and final tests can also further our understanding of the effect of effortful retrieval in learning. Perhaps higher levels of effort during the initial test produce more memory benefits (e.g., protection against forgetting) after a longer retention interval than the one used in this study.

Overall, the results of this study indicate that multiple-choice tests, besides a practical tool to measure learning (they are fast to complete and to grade), they are also a memory and self-efficacy booster regardless of the students’ age.
References


Appendix A

Reading Materials

Armadillos

Of the 20 varieties of armadillo, all but one live in Latin America. The familiar nine-banded armadillo is the only species that includes the United States in its range. Armadillo is a Spanish word meaning “little armored one” and refers to the bony plates that cover the back, head, legs, and tail of most of these odd looking creatures. Armadillos are the only living mammals that wear such shells.

Closely related to anteaters and sloths, armadillos generally have a pointy or shovel-shaped snout and small eyes. They vary widely in size and color, from the 6-inch-long (15-centimeter-long), salmon-colored pink fairy armadillo to the 5-foot-long (1.5-meter-long), dark-brown giant armadillos. Others have black, red, gray, or yellowish coloring. Contrary to popular belief, not all armadillos are able to encase themselves in their shells. In fact, only the three-banded armadillo can, curling its head and back feet and contorting its shell into a hard ball that confounds would-be predators.

Armadillos live in temperate and warm habitats, including rain forests, grasslands, and semi-deserts. Because of their low metabolic rate and lack of fat stores, cold is their enemy, and spates of intemperate weather can wipe out whole populations. Most species dig burrows and sleep prolifically, up to 16 hours per day, foraging in the early morning and evening for beetles, ants, termites, and other insects. They have very poor eyesight, and utilize their keen sense of smell to hunt. Strong legs and huge front claws are used for digging, and long, sticky tongues for extracting ants and termites from their tunnels. In addition to bugs, armadillos eat small vertebrates, plants, and some fruit, as well as the occasional carrion meal. Population numbers of nearly all species are threatened by habitat loss and over-hunting. Many cultures in the Americas consume armadillo flesh, which is said to resemble pork in its flavor and texture. Currently, only the nine-band population is expanding, and some species, including the pink fairy, are threatened.

Black Holes.

Black holes are the cold remnants of former stars, so dense that no matter—not even light—is able to escape their powerful gravitational pull. While most stars end up as white dwarfs or neutron stars, black holes are the last evolutionary stage in the lifetimes of enormous stars that had been at least 10 or 15 times as massive as our own sun. When giant stars reach the final stages of their lives they often detonate in cataclysms known as supernovae. Such an explosion scatters most of a star into the void of space but leaves behind a large "cold" remnant on which fusion no longer takes place.

In younger stars, nuclear fusion creates energy and a constant outward pressure that exists in balance with the inward pull of gravity caused by the star's own mass. But in the dead remnants of a massive supernova, no force opposes gravity—so the star begins to collapse in upon itself. With no force to check gravity, a budding black hole shrinks to zero volume—at which point it is infinitely dense. Even the light from such a star is unable to escape its immense gravitational pull. The star's own light becomes trapped in orbit, and the dark star becomes known as a black hole.

Black holes pull matter and even energy into themselves—but no more so than other stars or cosmic objects of similar mass. That means that a black hole with the mass of our own
sun would not "suck" objects into it any more than our own sun does with its own gravitational pull. Planets, light, and other matter must pass close to a black hole in order to be pulled into its grasp. When they reach a point of no return they are said to have entered the event horizon—the point from which any escape is impossible because it requires moving faster than the speed of light.

**Small But Powerful:** Black holes are small in size. A million-solar-mass hole, like that believed to be at the center of some galaxies, would have a radius of just about two million miles (three million kilometers)—only about four times the size of the sun. A black hole with a mass equal to that of the sun would have a two-mile (three-kilometer) radius. Because they are so small, distant, and dark, black holes cannot be directly observed. Yet scientists have confirmed their long-held suspicions that they exist. This is typically done by measuring mass in a region of the sky and looking for areas of large, dark mass.

Many black holes exist in binary star systems. These holes may continually pull mass from their neighboring star, growing the black hole and shrinking the other star, until the black hole is large and the companion star has completely vanished. Extremely large black holes may exist at the center of some galaxies—including our own Milky Way. These massive features may have the mass of 10 to 100 billion suns. They are similar to smaller black holes but grow to enormous size because there is so much matter in the center of the galaxy for them to add. Black holes can accreke limitless amounts of matter; they simply become even denser as their mass increases. Black holes capture the public's imagination and feature prominently in extremely theoretical concepts like wormholes. These "tunnels" could allow rapid travel through space and time—but there is no evidence that they exist.

**The Human Heart**

The heart is the body's engine room, responsible for pumping life-sustaining blood via a 60,000-mile-long (97,000-kilometer-long) network of vessels. The organ works ceaselessly, beating 100,000 times a day, 40 million times a year—in total clocking up three billion heartbeats over an average lifetime. It keeps the body freshly supplied with oxygen and nutrients, while clearing away harmful waste matter. The fetal heart evolves through several different stages inside the womb, first resembling a fish's heart, then a frog's, which has two chambers, then a snake's, with three, before finally adopting the four-chambered structure of the human heart.

About the size of its owner's clenched fist, the organ sits in the middle of the chest, behind the breastbone and between the lungs, in a moistened chamber that is protected all round by the rib cage. It's made up of a special kind of muscle (cardiac muscle) that works involuntarily, so we don't have to think about it. The heart speeds up or slows down automatically in response to nerve signals from the brain that tell it how much the body is being exerted. Normally the heart contracts and relaxes between 70 and 80 times per minute, each heartbeat filling the four chambers inside with a fresh round of blood. These cavities form two separate pumps on each side of the heart, which are divided by a wall of muscle called the septum. The upper chamber on each side is called the atrium. The left ventricle pumps most forcefully, which is why a person's heartbeat is felt more on the left side of the chest.
When the heart contracts, the chambers become smaller, forcing blood first out of the atria into the ventricles, then from each ventricle into a large blood vessel connected to the top of the heart. These vessels are the two main arteries. One of them, the pulmonary artery, takes blood to the lungs to receive oxygen. The other, the aorta, transports freshly oxygenated blood to the rest of the body. The vessels that bring blood to the heart are the veins. The two main veins that connect to the heart are called the vena cava.

**Blood Delivery:** Since the heart lies at the center of the blood delivery system, it is also central to life. Blood both supplies oxygen from the lungs to the other organs and tissues and removes carbon dioxide to the lungs, where the gas is breathed out. Blood also distributes nourishment from the digestive system and hormones from glands. Likewise our immune system cells travel in the bloodstream, seeking out infection, and blood takes the body's waste products to the kidneys and liver to be sorted out and trashed.

Given the heart's many essential functions, it seems wise to take care of it. Yet heart disease has risen steadily over the last century, especially in industrialized countries, due largely to changes in diet and lifestyle. It has become the leading cause of death for both men and women in the United States, claiming almost 700,000 lives a year, or 29 percent of the annual total. Worldwide, 7.2 million people die from heart disease every year.

**Tsunamis**

A tsunami is a series of ocean waves that sends surges of water, sometimes reaching heights of over 100 feet (30.5 meters), onto land. These walls of water can cause widespread destruction when they crash ashore. These awe-inspiring waves are typically caused by large, undersea earthquakes at tectonic plate boundaries. When the ocean floor at a plate boundary rises or falls suddenly it displaces the water above it and launches the rolling waves that will become a tsunami.

Most tsunamis, about 80 percent, happen within the Pacific Ocean’s “Ring of Fire,” a geologically active area where tectonic shifts make volcanoes and earthquakes common. Tsunamis may also be caused by underwater landslides or volcanic eruptions. They may even be launched, as they frequently were in Earth’s ancient past, by the impact of a large meteorite plunging into an ocean.

Tsunamis race across the sea at up to 500 miles (805 kilometers) an hour—about as fast as a jet airplane. At that pace they can cross the entire expanse of the Pacific Ocean in less than a day. And their long wavelengths mean they lose very little energy along the way.

In deep ocean, tsunami waves may appear only a foot or so high. But as they approach shoreline and enter shallower water they slow down and begin to grow in energy and height. The tops of the waves move faster than their bottoms do, which causes them to rise precipitously.

A tsunami’s trough, the low point beneath the wave’s crest, often reaches shore first. When it does, it produces a vacuum effect that sucks coastal water seaward and exposes harbor and sea floors. This retreating of sea water is an important warning sign of a tsunami, because the wave’s crest and its enormous volume of water typically hit shore five minutes or so later. Recognizing this phenomenon can save lives.

A tsunami is usually composed of a series of waves, called a wave train, so its destructive force may be compounded as successive waves reach shore. People experiencing a tsunami should remember that the danger may not have passed with the first wave and
should await official word that it is safe to return to vulnerable locations. Some tsunamis
do not appear on shore as massive breaking waves but instead resemble a quickly surging
tide that inundates coastal areas.
The best defense against any tsunami is early warning that allows people to seek higher
ground. The Pacific Tsunami Warning System, a coalition of 26 nations headquartered in
Hawaii, maintains a web of seismic equipment and water level gauges to identify
tsunamis at sea. Similar systems are proposed to protect coastal areas worldwide.
Appendix B

Testing materials

Armadillos
1. What kind of habitats can wipe out whole armadillo populations?
2. What does “armadillo” translate to?
3. Armadillos can be as big as:
4. Which type of armadillo is currently threatened?
5. Where do the majority of armadillo varieties live?
6. Armadillos mostly eat:
7. Which type or armadillo can contort itself into a hard ball to thwart predators?
8. What is the armadillo’s nose shaped like?
9. How many varieties of armadillos are there?
10. Armadillo meat resembles what kind of commonly eaten meat?

Black Holes
1. Compared to other cosmic objects of a similar mass, the gravitational force of a black hole is:
2. What can escape the gravitational pull of a black hole?
3. What is it called when a planet or other object reaches the point at which they can no longer avoid being pulled into a black hole?
4. What is the name of the fictional tunnel that allows one to time-travel through a black hole?
5. Large black holes exist at the center of:
6. What did black holes used to be?
7. In the dead remnants of the detonation before a black hole is formed, the matter begins to:
8. Before becoming black holes, the matter black holes come from are approximately how much bigger than our sun?
9. What is the volume of a black hole?
10. In the last stage before a black hole is formed, a detonation occurs, known as a(n):

The Human Heart
1. In addition to the liver, what organ helps sort and get rid of the body’s waste?
2. The heart supplies oxygen and removes __________ from the lungs.
3. What takes blood to the lungs and receives oxygen?
4. In the United States, approximately what percent of people die every year from heart disease?
5. Heart disease has risen over the last century, especially in __________.
6. The heart is divided by a wall of muscle called the __________.
7. Which part of the heart is the most powerful?
8. Approximately, how many times a day does a human heart beat?
9. How many chambers does the developed human heart have?
10. How long is the “network” of vessels that the human heart pumps blood through?
**Tsunamis**
1. What else (besides the main cause of tsunamis) can cause a tsunami?
2. What percentage of tsunamis occurs near the “Ring of Fire”?
3. In a tsunami, which part of the wave moves faster?
4. What is an important warning sign of a coming tsunami?
5. How high can tsunami waves reach on land?
6. What happens to a tsunami’s energy level as it moves?
7. Tsunamis are usually made up of a series of waves called a __________.
8. Where is the headquarters for the Pacific Tsunami System?
9. How fast can tsunami’s move across the ocean?
10. What typically causes a tsunami?
Appendix C

Self-efficacy for memory (adapted from Quinones, 1995).

Self-efficacy for memory before the initial session.
1. I feel confident in my ability to remember the information that I am about to read.
2. I think I can eventually reach a high level of memory for the information I am about to read.
3. I am sure I can effectively memorize the information that I am about to read in a relatively short period of time.
4. I don’t feel that my memory is as good as other people who are around my age.
5. On average, other people around my age probably have better memory than me.
6. I am fast at memorizing information, in comparison to other people who are around my age.
7. I am not sure I can ever reach a high level of memory for the readings, no matter how many times I read the information.
8. It would take me a long time to memorize the information in theses readings.
9. I am not confident that I can successfully remember the readings.
10. I doubt that my memory would be very adequate to remember the readings.

Self-efficacy for memory after each test during the initial session.
1. I feel confident in my ability to remember the information in the article during the final test.
2. I am sure I can remember the information in the article for the final test.
3. When compared to other people who are around my age, I don’t feel that I am as capable of remembering the information in the article for the final test.
4. On average, other people around my age are probably much more capable than me at remembering the information in the article for the final test.
5. I doubt that my ability to answer the test correctly will be adequate.

Self-efficacy for memory after completing the final test (second session).
1. I feel confident in my ability to remember information in the future.
2. I think I can remember much of the information in the readings.
3. I am sure I can remember the information in the readings in the future.
4. I don’t feel I am as capable of remembering the information as other people who are around my age.
5. On the average, other people around my age are probably much more capable of remembering the information after a delay.
6. I am good at remembering information after a long delay, in comparison to other people around my age.
7. I am not sure I could ever remember the information on the readings, no matter how many times I have read it.
8. It would take many times of me reading the information in order to effectively answer a test after a delay.
9. I am not confident that I can retain successfully the information I read.
10. I doubt that memory for this information will be adequate after a delay.
Appendix D

Demographic information questionnaire

Please enter your age:

Please select the highest level of education achieved:

Elementary
Junior high
High school
Some college
Associates degree
Bachelor degree
Some graduate work
Graduate degree