ABSTRACT

LUONG, MARY G. Refining the Retrieval Effort Theory of the Testing Effect as it Relates to Feedback. (Under the direction of Dr. Christopher B. Mayhorn and Dr. Brad Mehlenbacher).

Testing effect is a cognitive phenomenon wherein the very act of taking a test promotes longer retention of studied information compared to restudying the material. The retrieval effort theory broadly explains that the mechanism behind the testing effect is the cognitive effort expended in the retrieval of answers from memory such that more effort should result in longer retention than less effort. Research has shown that retention is further improved over testing alone when feedback is given during one or more testing events preceding a final test of retention. However, little research to date has been conducted to determine if using questions as facilitative feedback has an effect on retrieval effort and retention. The current study investigated whether there is an additive benefit to learning of leveraging the testing effect through feedback that is, itself, a test. Findings demonstrated that Facilitative Feedback Questions (FFQs) indeed offer an additive benefit however the implication to the retrieval effort theory lies not in retrieval effort, per se, but rather in successful retrieval effort.
Refining the Retrieval Effort Theory of the Testing Effect as it Relates to Feedback

by
Mary G. Luong

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

__________________________________________
Dr. Christopher B. Mayhorn
Committee Co-Chair

__________________________________________
Dr. Brad Mehlenbacher
Committee Co-Chair

__________________________________________
Dr. Douglas J. Gillan

__________________________________________
Dr. Kristy E. Boyer
DEDICATION

This work is dedicated to friends, colleagues, mentors, and most especially loved ones who encouraged and bolstered me throughout its highs and lows. To each and all, I am grateful for the passion you fueled and the strength you built so I could cross this finish line and give back to those who pursue knowledge tirelessly.
BIOGRAPHY

I spent many, many years as a computer geek before I went back to college. I earned my B.A. in 2008 at Guilford College, a wonderful, Quaker, liberal arts college in Greensboro, North Carolina. From there, I pursued my graduate education at North Carolina State University. Among the wolf pack I found students, staff, and faculty that challenged me in every way imaginable. After graduate school, I am looking forward to blending the skills I acquired in industry and the knowledge I gained in academia.
ACKNOWLEDGMENTS

I’d like to thank my co-chairs, Dr. Christopher Mayhorn and Dr. Brad Mehlenbacher. Your perspectives and support made the culmination of my graduate school experience one I will look back on with immense gratitude and a sense of humor.

Dr. Gillan and Dr. Boyer’s curiosity about my research provided fuel when my own curiosity waned. For that investment, I would like to extend my thanks.

A special thanks goes out to Jennifer London who supported me through even the darkest of my dissertation writing hours!
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Retrieval Effort Theory of the Testing Effect</td>
<td>2</td>
</tr>
<tr>
<td>Retrieval Difficulty</td>
<td>4</td>
</tr>
<tr>
<td>Question Format</td>
<td>5</td>
</tr>
<tr>
<td>Feedback</td>
<td>7</td>
</tr>
<tr>
<td>FFQs</td>
<td>10</td>
</tr>
<tr>
<td>Study Overview and Research Questions</td>
<td>10</td>
</tr>
<tr>
<td>METHOD</td>
<td>12</td>
</tr>
<tr>
<td>Design</td>
<td>12</td>
</tr>
<tr>
<td>Participants</td>
<td>12</td>
</tr>
<tr>
<td>Materials</td>
<td>14</td>
</tr>
<tr>
<td>Procedure</td>
<td>15</td>
</tr>
<tr>
<td>RESULTS</td>
<td>18</td>
</tr>
<tr>
<td>Preliminary Analyses</td>
<td>18</td>
</tr>
<tr>
<td>Tests of Research Questions</td>
<td>19</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>24</td>
</tr>
<tr>
<td>Summary of Results</td>
<td>24</td>
</tr>
<tr>
<td>Implications</td>
<td>26</td>
</tr>
<tr>
<td>Limitations</td>
<td>27</td>
</tr>
<tr>
<td>Future Research</td>
<td>28</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>29</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>36</td>
</tr>
<tr>
<td>Appendix A: Mechanical Turk Informed Consent</td>
<td>37</td>
</tr>
<tr>
<td>Appendix B: Undergraduate Psychology Course Informed Consent</td>
<td>39</td>
</tr>
<tr>
<td>Appendix C: Demographics Survey</td>
<td>41</td>
</tr>
<tr>
<td>Appendix D: Devil’s Hole Fact Sheet</td>
<td>42</td>
</tr>
<tr>
<td>Appendix E: Geology Questions</td>
<td>44</td>
</tr>
<tr>
<td>Appendix F: Debriefing Document</td>
<td>49</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Descriptive Statistics of Test and WMC Scores by Type of Feedback ..................16
Table 2. Descriptive Statistics by Amount of Feedback Seen on Intervening-Test Among Participants Who Received Feedback.................................................................20
Table 3. Descriptive Statistics by Level of Retrieval Success on Intervening-Test for All Participants........................................................................................................21
LIST OF FIGURES

Figure 1. Study Design Flowchart. ........................................................................................................... 12

Figure 2. Data Collection Flowchart ........................................................................................................ 15

Figure 3. Mean Retention-Test Scores by Type of Feedback ................................................................. 20
INTRODUCTION

Your psychology professor believes in the power of the testing to improve learning. At the beginning of each class she asks questions about the day’s reading assignment. Today your professor’s question hangs in the air while she waits for your answer. You pause, trying to retrieve the answer from memory, and give a response. Then professor says, “Let me rephrase the question.” Implicit in the rephrased question is that the first answer was either incorrect or did not contain all the information that the professor was seeking. In other words, she offers you feedback.

Feedback in the course of testing is often critical to improving learning outcomes (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Epstein et al., 2002). Simply finding out that an answer is correct or incorrect is better for retention than getting no feedback (Hanna, 1976). Research on feedback, however, has largely been limited to static text that is shown to the learner. Rephrasing a question makes feedback interactive and its value to learning can be seen when viewed in the context of the testing effect. The testing effect is a cognitive phenomenon wherein the very act of taking a test promotes longer retention of studied information compared to restudying the material. Despite the power of testing alone (Cull, 2000; Kuo & Hirshman, 1996), many researchers have integrated feedback to investigate its impact on the testing effect (Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007; Kulhavy & Stock, 1989; McDaniel & Fisher, 1991; Pashler, Cepeda, Wixted, & Rohrer, 2005).
Researchers have found that retention is further improved over testing alone when feedback is given during one or more testing events preceding a final test of retention (Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006b). Among testing-effect studies, these gains have been experimentally observed using two kinds of feedback: informing participants whether their response was correct or not and displaying the correct answer to the participant. Considering this relationship between testing and feedback, and that testing has been shown to directly produce learning gains, the present study investigated whether there is an additive benefit to learning of leveraging the testing effect through feedback that is, itself, a test (e.g., a rephrased question). The aim of the research was to refine the retrieval effort theory of the testing effect, as it relates to feedback. The following sections will introduce the retrieval effort theory, present an overview of retrieval difficulty, summarize characteristics of feedback in the general context of retrieval difficulty, and offer an explanation of how giving rephrased questions as feedback may serve to refine the retrieval effort theory of the testing effect, as it relates to feedback.

**Retrieval Effort Theory of the Testing Effect**

Retrieval practice, or testing, results in the cognitive phenomenon known as the testing effect, which describes an improvement in retention of studied materials when studied materials are tested versus re-studied (Hogan & Kintsch, 1971; Jacoby, 1978; Nungester & Duchastel, 1982a). This robust phenomenon has been found using images (Wheeler & Roediger, 1992), maps (Carpenter & Pashler, 2007), and text, to name a few. Text-based stimuli have ranged from word lists (Karpicke & Roediger, 2007a, 2007b; Tulving, 1967) to educationally-relevant materials such as textbook passages and topic-specific articles.
(Agarwal, Karpicke, Kang, Roediger, McDermott, 2007; Duchastel, 1981; Glover, 1989; Roediger & Karpicke, 2006a); the latter being of interest to the present study. The retrieval effort theory broadly explains that the mechanism behind the testing effect is the cognitive effort expended in the retrieval of answers from memory (Roediger & Karpicke, 2006a). More importantly, the theory attempts to predict variations in the magnitude of the testing effect such that more cognitive effort employed during testing results in longer retention of the tested material (i.e., higher final retention-test scores).

A wealth of testing effect research has shown that effortful retrieval of information from memory improves retention of that knowledge more than when it is easily recalled. Gardiner, Craik, and Bleasdale (1973) used definitions of uncommon words and captured retrieval latency as the measure of retrieval effort. Over the course of 50 trials, the experimenters presented participants with “definitions of 50 [uncommon] words whose frequency count was less than 1 per million” (p. 213). After each definition, participants tried to retrieve the target word while the researchers manually recorded the time until a target was retrieved (i.e., retrieval latency). The next definition was presented after successful word retrieval or 90 seconds had elapsed; target words were not given if the participant failed to recall them. After the last trial, participants waited one minute and were asked to recall all 50 words in a free recall retention test. Words with the longest retrieval latency had the lowest total recall frequency in the definition session but they had the highest recall probabilities on the final retention test. The researchers concluded that words that were initially difficult to retrieve, as measured by longer retrieval latency, resulted in better recall on a final retention test than words that were easily retrieved, as measured by shorter retrieval latency. These
results lend empirical support to the retrieval effort theory; however, they only tell part of the story. By using equally uncommon words, there was no variation in retrieval difficulty.

**Retrieval Difficulty**

Retrieval difficulty follows from Bjork’s (1994, 1999) framework of desirable difficulties. Desirable difficulties are the basis of pedagogical techniques for improving long-term retention that (may) negatively impact short term learning. The term is used to describe conditions under which learning (or more specifically, encoding) is effortful (Schmidt & Bjork, 1992) up to a surmountable level. Retrieval difficulty refers to elements of the task whereas retrieval effort is related to the individual’s behavior while engaged in the task; therefore, retrieval difficulty can be manipulated to influence retrieval effort. Learning occurs when individuals expend more retrieval effort while engage in difficult memory tasks that result in successful retrieval. Support for desirable difficulties and retrieval effort has been presented by Pyc and Rawson (2009) who found that participants in the high difficulty but successful retrieval group had higher final test scores (i.e., long-term retention) than those in the low difficulty successful retrieval group. According to Roediger and Karpicke (2006a), testing is an effortful memory task that has desirable difficulty because tests can have varying levels of retrieval difficulty that affect retrieval effort.

There are a number of task elements that can be manipulated to vary retrieval difficulty of a test. In an excellent review by Roediger and Karpicke (2006b), studies have demonstrated the testing effect’s robustness using a breadth of task elements that vary retrieval difficulty, including “the materials to be learned, the format of [questions on] the initial and final tests, whether or not subjects receive feedback on the first test, the time
interval between studying and initial testing, and the retention interval before the final test, to name but a few” (p. 182). Of particular interest to the present study were question format and type of feedback.

**Question Format.** Multiple choice questions generally yield high scores but are considered to be at the low end of the retrieval difficulty spectrum because the correct response is guaranteed to be present in the list of choices (Shute, 2008). The retrieval effort employed in answering multiple choice questions, however, can vary widely depending upon their psychometric quality. The proposed study will exclude multiple choice questions for three reasons. First, Roediger and Marsh (2005) found that lures (i.e., distractor choices) in multiple choice questions may lead to acquiring false knowledge. Second, lures can also cause a learner to incorrectly apply existing knowledge (Roediger & Marsh, 2005) thereby diluting the measure of an individual’s true subject knowledge. Third, multiple choice questions have low fidelity with real world interactions. For example, it is an unlikely scenario that one’s manager would ask a question followed by a series of response options. It is more likely that the manager is inquiring to be informed rather than to assess the employee’s declarative knowledge.

On the high end of the retrieval difficulty and retrieval effort spectrums are recall questions because the correct response must be produced or recalled from memory (Shute, 2008). The recall category is further divided into free- and cued-recall subcategories with free-recall demanding more retrieval effort than cued-recall. Open-ended questions can fall in either the free- or cued-recall category. When the question stem simply instructs the learner to recall from memory as much of the studied material as possible, the question is placed in
the free-recall category. Alternatively if the question stem provide cues (e.g., topic sentences) followed by the above instructions it would be placed in the cued-recall category. The current study excluded open-ended questions due to concerns about accuracy of automated scoring.

Short-answer and completion (e.g., fill-in-the-blank) questions are generally categorized as cued-recall because of the focused nature of the question stems. Osterlind (1998) advised that the “correct use of short-answer and sentence-completion questions is in assessing areas that require a great deal of factual recall” (p. 238). To test this factual knowledge the stems for short-answer and completion questions are necessarily specific, which can have a cuing effect. Osterlind (1998) gave the following two stems to illustrate the importance of specificity in these types of questions:

A) “Rain is produced by ______________ clouds” (p. 239).

B) “During thunderstorms, rain is produced by ______________ clouds” (p. 239).

The thunderstorm qualifier, which acts as a cue, is important for three reasons. First, A is not specific enough to allow for a one- or two-word response because rain can be produced by several types of clouds. Second, a guess might be the correct response to A but not to B. Third, there is only one correct answer to B.

Although research has shown that free-recall, open-ended questions demand more retrieval effort and result in higher long-term retention than cued-recall, short-answer or completion questions (Glover, 1989), there are learning gains achieved by the latter (Carpenter & DeLosh, 2006; Hinze & Wiley, 2011; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985; Nungester & Duchastel, 1982b).
Feedback. Black and William (1998) described feedback as having two functions: directive and facilitative. Directive feedback is specific and gives learners the details of a problem (e.g., an erroneous response to a test question) and its solution. This directing of the learner often requires little to no retrieval and thus falls on the low end of the retrieval difficulty spectrum. Facilitative feedback, on the other hand, is more general and guides learners to self-identification of the problem and solution (Gilman, 1969). As such, facilitative feedback tends toward the high end of retrieval difficulty.

In an in-depth review, Shute (2008) noted that “… good feedback can significantly improve learning processes and outcomes, *if delivered correctly* [italics added]” (p. 154). Studies of feedback, generally, and facilitative feedback, specifically, have attempted to address three issues related to delivery and, in turn, retrieval difficulty: specificity (verification versus elaboration), complexity (feedback types), and timing (immediate versus delayed). The following three sections will review the findings from the feedback literature related to these issues.

*Feedback specificity.* Feedback specificity is defined as the level of information contained in feedback messages (Goodman, Wood, & Hendrickx, 2004). Specificity is related to whether feedback is directive or facilitative because if feedback contains all the information needed to solve a problem, it is necessarily directive. As feedback becomes less specific, the user is required to make more connections between the information that is given and the knowledge she possesses to produce a correct response. This reduction in specificity and subsequent increase in the necessary connections results in feedback becoming more cognitively demanding and, compared to directive feedback, more difficult to produce a
correct response. In other words, retrieval difficulty increases as specificity decreases. There is no level of specificity recognized as optimal, but research has shown that feedback that gives enough information to improve a response is more effective and that knowledge of results (i.e., correct/incorrect) feedback is less effective as well as frustrating (Bangert-Drowns et al., 1991; Kulhavy & Stock, 1989; Phye & Sanders, 1994). Some researchers have suggested that specificity may have an inverted U-shaped relationship with learning (Hogarth, Gibbs, McKinzie, & Marquis, 1991; Irion, 1969; Salmoni, Schmidt, & Walter, 1984). Unfortunately, the support for this position has not been consistent, and what support for the inverted-U hypothesis that has been obtained is not consistent with a specificity or precision explanation (Kluger & Denisi, 1996).

*Feedback complexity.* Feedback complexity is often discussed together with feedback length because, in most cases, they represent the same challenges. On the spectrum of complexity, a great deal of information can lead to information overload on one end and a dearth of information to confusion on the other. Some examples of feedback, in order of complexity and retrieval difficulty, are: correct/incorrect, error flagging, elaborated, and topic contingent (see Table 1, Shute, 2008). Whereas correct/incorrect feedback provides the learner one piece of information per item, error flagging can give multiple pieces of information as well as locational information. Similarly, elaborated feedback refers to the provision of some general explanation of why the response was incorrect or provides a general framework for coming to the correct responses, and topic contingent feedback provides information from that item’s content domain. Topic contingent feedback can be likened to re-teaching a chapter for a relevant item. According to Shute (2008), even the
simplest feedback should provide information on whether an answer is incorrect and give some clue as to how the answer could be improved. Meta-analytic research by Sleeman, Kelly, Martinak, Ward, and Moore (1989) revealed that there were too many conflicting findings to reach a consensus regarding optimal complexity. It seems likely that situational moderators, such as degree of specificity of the test itself, may play a role in determining what level of feedback complexity is most congruent with learning.

Feedback timing. Feedback timing refers to when the feedback is provided to the learner. In most university settings, students receive feedback at the test level a few days after finishing a paper-based exam; this is known as delayed feedback. With online testing immediate feedback can be provided after a student submits a response. There are a number of contradictory findings regarding which feedback timing is superior for learning, and each side has supporters (Kulik & Kulik, 1988). Proponents of delayed feedback cite the interference-perseveration hypothesis, which is that learners forget the flawed notions on which they initially perseverated and thus they do not interfere with retention when feedback is provided after a delay (e.g., Kulhavy & Anderson, 1972; Schroth, 1992). Proponents of immediate feedback argue that it is more efficient to correct as the mistake is made (e.g., Anderson, Magill, & Sekiya, 2001). Although the majority of research is too contradictory to draw conclusions from, there have been some consistent findings that immediate feedback is more effective for programming and math than is delayed feedback (Corbett & Anderson, 2001). However, whether this is due to the content or the delivery mechanism is not clear.
FFQs

Feedback that is a rephrased question is itself a test. A rephrased question has the potential to act as facilitative feedback because rephrasing of an initial question is an example of elaboration. A search of the feedback literature for the use of questions as facilitative feedback yielded a dearth of results. The potential for Facilitative Feedback Questions (FFQs) to refine the retrieval effort theory comes from its basis in task elements that, conceptually, create high task difficulty and high retrieval effort.

The evidence-based design of FFQs was derived from the following findings in both the testing effect and feedback literature. First, testing that requires more retrieval effort results in longer retention of tested material compared to low retrieval effort testing. Second, cued-recall questions, though less powerful than free-recall questions, still demonstrate improvements in long-term retention for tested material. Third, providing facilitative feedback that guides learners to self-correct erroneous responses improves learning compared to simply giving the learner the correct response. Fourth, topic-contingent feedback provides clues as to the correct answer which aids in learning. And finally, immediate feedback appears to be more effective than delayed feedback to improve learning of fact-based material (e.g., a geology article). Thus, FFQ design incorporates the characteristics of testing and feedback that are highly effective in aiding retention.

Study Overview and Research Questions

The aim of the present research was to refine the retrieval effort theory of the testing effect, as it relates to feedback. To that end, the study investigated whether there is an additive benefit to learning of leveraging the testing effect through feedback that is, itself, a
test. Due to the retrieval difficulty inherent in recall questions, retrieval effort was manipulated to be high by creating a test comprised of short-answer questions. If, as speculated by Kang, McDermott, and Roediger (2007), increased retrieval difficulty leads to improved feedback encoding, then keeping retrieval difficulty high across all experimental conditions should increase the sensitivity of feedback to produce a measurable effect on retention. In addition, comparing feedback that requires retrieval effort to feedback that does not require any retrieval effort should also increase the potential to find a measurable effect of FFQs on retention.

A secondary purpose of the present research was to acknowledge the possible moderating effect of working memory capacity (WMC) on the testing effect as described by Tse and Pu (2012). In the present research, WMC was included to determine whether it would moderate any efficacy found in FFQs. Because working memory is part of short-term memory, its effect was measured on the intervening-test and not the retention-test, which were taken immediately after reading the geology article and after a one-week retention interval, respectively.

Thus, the following research questions were investigated:

Research Question 1: Do FFQs offer an additive benefit to learning by leveraging the testing effect through feedback that is, itself, a test?

Research Question 2: Does working memory capacity (WMC) moderate the strength or directionality of the relationship between FFQs and intervening-test score?
METHOD

Design

The study used a pre-test, study (geology article), intervening-test (retrieval practice), retention-test design with a one-week retention interval between the intervening-test and the retention-test (see Figure 1).

![Study Design Flowchart](image)

Figure 1. Study Design Flowchart.

Participants

A total of 236 participants started the study but only 195 returned after the one-week retention interval to complete the study, resulting in an attrition rate of 17%. Of the 195 participants who completed the entire study 18 were dropped from the analyses because they failed to follow instructions; e.g., did not provide their code upon returning from the retention interval or gave inappropriate answers. A total of 177 participants are included in the analyses. Twenty-two of the participants were recruited through an undergraduate psychology course at North Carolina State University and were awarded extra credit in the
The remaining 155 participants were recruited through Amazon’s Mechanical Turk online crowdsourcing tool and paid $2.00 for their participation. To minimize reading comprehension confounds, only U.S.-based Mechanical Turk “Workers” were allowed to accept the human-intelligence task (HIT) posted on Mechanical Turk. Behrend, Sharek, Meade, and Wiebe (2011) found that Mechanical Turk is an effective tool for recruiting online behavioral research participants.

The sample was 59% female and the ages were highly varied with a mean of 34 and a standard deviation of 13 years of age. The youngest participant was 19 years of age and the oldest 67. The largest category of educational attainment was completion of an undergraduate degree (41%). Additionally, 39% of participants completed a portion of an undergraduate degree, 11% completed a graduate degree, and 8% completed high school. Only one participant (<1%) did not complete high school. The majority of the sample reported that they were Caucasian (85%), and representation of other ethnicities was split amongst the rest: African-American (5%), Asian (3%), Hispanic (3%), Other (3%), Native American (<1%), and Pacific Islander (<1%). For all but three participants, English was the primary language.

Participants were asked about their confidence in their ability to pass a geology test (with a grade of 70% or higher) and their level of academic attainment with geology in order to help explain differences due to interest or prior experience with the topic. The three-point confidence scale was from 1: Very Confident to 3: Not Confident and the mean score was 2.22 (SD = .61). In terms of geology-related experience, 94% answered none/high school, 2% had an undergraduate degree in geology, and 4% were interested in geology as a hobby.
Materials

The symmetry span test (Unsworth, Redick, Heitz, Broadway, & Engle, 2009; Conway, Kane, Bunting, Hambrick, Wilhelm, Engle, 2005) is a measure of working memory capacity and an automated version (Redick, Broadway, Meier, Kuriakose, Unsworth, Kane, & Engle, 2012) was administered by Georgia Institute of Technology’s Attention and Working Memory Lab (http://psychology.gatech.edu/rengelelab) with the help of Dr. Jeffrey Foster’s Cog Lab website. Cog Lab was also used to administer the informed consents (see Appendices A and B). Qualtrics was used for the demographics survey (see Appendix C), to display a re-printed earthquake science article (see Appendix D), and administer a 17-question short-answer test (see Appendix E) during three separate testing events: pre-test, intervening-test, and retention-test. Qualtrics was also used to display the debrief document at the end of the experiment (see Appendix F).

The Devil’s Hole Fact Sheet was prepared by the U.S. Geological Survey as a primer intended for a broad audience of lay readers. Geology was chosen because it is a fact-based discipline and most people lack a deep understanding of geological processes and terminology. Due to the highly technical language content in the article, readability tests were conducted. The text’s Flesch-Kindcaaid Grade Level was 15.4 and Flesch Reading Ease test score was 29, which is very difficult to read. These characteristics made geology an ideal topic to measure declarative knowledge learning where the subject matter is new but not entirely foreign. All images and references to them were removed from the original articles to ensure it could be used as part of a verbal task. Sixteen of the 17 short-answer questions were written to test participants’ knowledge of the material presented in the geology article.
The seventeenth question was a check of the participant’s attention. In addition, each of the 16 short-answer, geology questions had a corresponding Facilitative Feedback Question (FFQ). The 16 FFQs (see Appendix E) were crafted by reformatting each short-answer questions into a fill-in-the-blank question and adding elaborative information to aid in self-correcting erroneous response to that short-answer question. All elaborative information was taken directly from the Devil’s Hole Fact Sheet.

**Procedure**

Participants were asked to digitally sign an electronic informed consent, which included a notice that this was a two-part study with part two to be completed one week after part one. Following the consent form, an automated symmetry span test was administered to measure working memory capacity and then demographic data were collected. Figure 2 illustrates the flow of the remainder of the procedure.
Figure 2. Data Collection Flowchart.

For the pre-test, participants were given the 17 short-answer questions, presented in random order. Qualtric’s Display Logic was used to award one point for each correct answer (see Appendix E for values accepted as correct). Following the pre-test, participants were randomly assigned to one of three experimental groups, based on feedback type – NO FEEDBACK, CORRECT ANSWER, or FFQ (Facilitative Feedback Question) – and taken to their respective instruction pages.

For the CORRECT ANSWER group, instructions included the following statement: “If you get an answer wrong, the correct answer will be shown to you.” For the FFQ group, instructions included the following statements: “If your first answer is wrong, you will be given a second chance and shown a fill-in-the-blank question. If your second answer is
wrong, the correct answer will be shown to you.” Both the CORRECT ANSWER and FFQ groups were given a multiple choice question asking what they would see if their answer (to the short-answer question) was wrong. All CORRECT ANSWER and FFQ participants answered the multiple choice question correctly.

Participants then proceeded to the geology article. After reading the article, the 17 short-answer questions were given again as the intervening-test, presented in a new random order. Qualtric’s Display Logic was used to award one point for each correct answer. The intervening-test contained the experimental manipulation. For the NO FEEDBACK group, participants were immediately taken to the next question after submitting each answer. For the CORRECT ANSWER group, an incorrect response was followed by a re-presentation of the question with the correct answer below it as feedback (see Appendix E). After acknowledging the feedback, participants were shown the next question. For the FFQ group, an incorrect response was immediately followed by the current question’s corresponding FFQ as feedback. In this scenario, participants were given a second opportunity to earn a point by answer the FFQ correctly. An incorrect answer to the FFQ was followed by the correct answer, which was formatted to match the CORRECT ANSWER group’s feedback (see Appendix E). Following a correct answer to the FFQ or acknowledging the correct answer feedback, participants were shown the next question.

After completing the intervening test, Mechanical Turk participants were given an initial completion code to enter into Mechanical Turk. This completion code signaled the end of part one. Participants entered the completion code into their Mechanical Turk HIT and after verification of the completion code participants were paid $0.50. The NCSU
undergraduates were also given an initial completion code along with a URL to part two and instructions to email themselves the code and URL. The students were told that URL for part two would not be open until 7 days had elapsed.

Following the one-week retention interval, Mechanical Turk participants were emailed (anonymously via the Mechanical Turk system) instructions along with a website link to begin part two of the study. NCSU undergraduates were emailed, by their instructor, a notice that the part two URL was active. After submitting part one’s completion code, all returning participants were given the 17 short-answer questions again as the retention-test, presented in a newly randomized order with no feedback. Qualtric’s Display Logic was used to award one point for each correct answer. After completing the retention-test, participants received their debriefing document. The remaining $1.50 was paid to each Mechanical Turk participant using Mechanical Turk’s bonus payment feature. The NSCU undergraduates each received their extra credit.

RESULTS

Preliminary Analyses

Descriptive statistics of test scores and working memory capacity (WMC) scores are presented in Table 1 by type of feedback. Pre-test score was non-normally distributed, with skewness of 3.82 ($SE = .18$) and kurtosis of 15.70 ($SE = .36$); transformation could not correct the non-normal distribution. Intervening-test and retention-test scores were normally distributed, as assessed by z-scores using a critical value of $|2.58|$ at the $p = .01$ level of significance.
A one-way ANOVA was conducted to determine if WMC score was different across the three feedback groups. Levene's test of homogeneity of variances was non-significant, \( p = .17 \). The differences in WMC score between these groups was not statistically significant, \( F(2, 173) = 1.80, p = .17 \).

Table 1. *Descriptive Statistics of Test and WMC Scores by Type of Feedback*

<table>
<thead>
<tr>
<th></th>
<th>NO FEEDBACK (N=62)</th>
<th>CORRECT ANSWER (N=54)</th>
<th>FFQ (N=61)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M )</td>
<td>( SD )</td>
<td>( M )</td>
</tr>
<tr>
<td>Pre-Test</td>
<td>1.19</td>
<td>2.35</td>
<td>1.22</td>
</tr>
<tr>
<td>Intervening-Test</td>
<td>6.79</td>
<td>3.55</td>
<td>5.41</td>
</tr>
<tr>
<td>Retention-Test</td>
<td>4.66</td>
<td>3.12</td>
<td>5.20</td>
</tr>
<tr>
<td>WMC</td>
<td>28.06</td>
<td>9.07</td>
<td>25.54</td>
</tr>
</tbody>
</table>

*Notes: \( \text{N} = 60 \). The test scores are out of 15 possible points; 1 question was an attention check and 1 question was thrown out due to a programming error by the researcher (see Appendix E). The WMC scores are out of 42 possible points.*

Tests of Research Questions

*Research Question 1:* Do FFQs offer an additive benefit to learning by leveraging the testing effect through feedback that is, itself, a test?

A one-way ANOVA was conducted to determine if retention score was different across different types of feedback. Levene's test of homogeneity of variances was non-significant, \( p = .25 \). While retention-test score increased from the NO FEEDBACK, to CORRECT ANSWER, to FFQ feedback groups (see Table 1 and Figure 3), the differences between these feedback groups was not statistically significant, \( F(2,174) = .87, p = .42 \). Not
only was FFQ not significantly different from CORRECT ANSWER but, inconsistent with the literature, neither feedback group was significantly better than NO FEEDBACK.

![Figure 3](image-url)

*Figure 3. Mean Retention-Test Scores by Type of Feedback.*

Two alternate interpretations of the data were investigated. The first was based on the amount of feedback seen. The second was based on the desirable difficulties framework of difficult but *successful* retrieval practice. To investigate amount of feedback seen, feedback counts (Mdn = 11) for participants in CORRECT ANSWER and FFQ groups were recoded into FEEDBACK SEEN with two levels: LESS and MORE. Counts of feedback seen from 0 to 11 were recoded as LESS FEEDBACK SEEN and counts from 12 to 27 became MORE FEEDBACK SEEN. An independent-samples t-test was run to determine if there were significant differences in retention-test scores between participants who saw LESS OR MORE FEEDBACK on the intervening-test (see Table 2 for descriptive statistics). Levene's
test of homogeneity of variances was significant, $p = .01$. The retention-test score was significantly higher for the LESS FEEDBACK SEEN group than for the MORE FEEDBACK SEEN group, $t(108.92) = 4.53, p < .01$. In other words, participants who saw more feedback did not have better retention-test scores than those who saw less feedback.

Table 2. Descriptive Statistics by Amount of Feedback Seen on Intervening-Test Among Participants Who Received Feedback

<table>
<thead>
<tr>
<th></th>
<th>LESS FEEDBACK SEEN (N = 58)</th>
<th>MORE FEEDBACK SEEN (N = 57)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>Feedback Count</td>
<td>8.05</td>
<td>2.05</td>
</tr>
<tr>
<td>Retention-Test</td>
<td>6.76</td>
<td>3.71</td>
</tr>
</tbody>
</table>

To investigate successful retrieval practice, the intervening-test scores across all participants (Mdn = 7) were recoded into RETRIEVAL SUCCESS with two levels: LOW and HIGH. Scores from 0 to 6 on the intervening-test were recoded as LOW RETRIEVAL SUCCESS and scores from 7 to 15 became HIGH RETRIEVAL SUCCESS. An independent-samples t-test was run to determine if there were differences in retention-test scores between participants who had LOW and HIGH RETRIEVAL SUCCESS on the intervening-test (see Table 3 for descriptive statistics). Levene's test of homogeneity of variances was significant, $p < .01$. The retention-test score was significantly lower for participants with LOW RETRIEVAL SUCCESS on the intervening-test than those with HIGH RETRIEVAL SUCCESS, $t(151.51) = -6.82, p < .01$. Consistent with the literature,
these results show that participants, who had more successful retrievals during retrieval practice, had higher retention-test scores than those who had fewer successful retrievals.

Table 3. Descriptive Statistics by Level of Retrieval Success on Intervening-Test for All Participants

<table>
<thead>
<tr>
<th></th>
<th>LOW RETRIEVAL SUCCESS (N = 88)</th>
<th>HIGH RETRIEVAL SUCCESS (N = 89)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Intervening-Test</td>
<td>3.91</td>
<td>1.86</td>
</tr>
<tr>
<td>Retention-Test</td>
<td>3.51</td>
<td>2.39</td>
</tr>
</tbody>
</table>

To better understand the RETRIEVAL SUCCESS results, a one-way ANOVA was conducted to determine if intervening-test score was different between the three types of feedback. Levene's test of homogeneity of variances was non-significant, $p = .06$. Intervening-test score was significantly different between different feedback groups, $F(2,174) = 11.48, p < .001, \omega^2 = .11$. FFQ had the highest intervening-test score, followed by NO FEEDBACK and then CORRECT ANSWER (see Table 1). Tukey post-hoc analysis revealed that the mean increase from NO FEEDBACK to FFQ (1.70, 95% CI [.23, 3.18]) was statistically significant ($p = .02$), as was the increase from CORRECT ANSWER to FFQ (3.08, 95% CI [1.56, 5.61], $p < .001$). There was no statistically significant difference between NO FEEDBACK and CORRECT ANSWER.

Given that intervening-test scores were significantly highest in the FFQ group and that retention-test scores were significantly higher in the HIGH RETRIEVAL SUCCESS
group than in the LOW RETRIEVAL SUCCESS group, FFQ data was selected for further analysis. Two Pearson's product-moment correlations were run to assess the relationship between successful retrieval on the intervening-test and correctly answering that same question on the retention-test in FFQ participants. In other words, is there a relationship between successfully retrieving the answer to question A on the intervening-test and successfully answering question A on the retention-test?

There was a large positive correlation between number of correctly answered FFQs and number of matching retention-test questions answered correctly, $r(54) = .62, p < .001$, with number of FFQs answered correctly explaining 38% of the variation in matching retention-test items answered correctly. Similarly, there was a large positive correlation between number of correctly answered short-answer questions (in the FFQ group) and number of matching retention-test questions answered correctly, $r(54) = .63, p < .001$, with number of short-answer questions answered correctly explaining 40% of the variation in matching retention-test items answered correctly.

*Research Question 2:* Does working memory capacity (WMC) moderate the strength or directionality of the relationship between type of feedback and intervening-test score?

A moderated multiple regression was conducted with intervening-test score as the criterion. In Step 1, dummy-coded feedback conditions, with FFQ as the baseline group, were entered as main effect predictors. In Step 2, the interaction terms were entered into the model but neither interaction term was significant, $\Delta R^2 = .01, \Delta F(1,170) = .61, p = .55$. In other words, WMC did not moderate the relationship between type of feedback and intervening-test score.
To further explore WMC, two independent samples t-tests were conducted with FEEDBACK SEEN and RETRIEVAL SUCCESS to determine if there were significant differences in WMC between participants in each of those recoded groups. For FEEDBACK SEEN, Levene's test of homogeneity of variances was non-significant, $p > .05$. There was no significant difference in WMC between participants who saw LESS FEEDBACK ($M = 27.47$, $SD = 8.10$) and those who saw MORE FEEDBACK ($M = 26.19$, $SD = 6.89$) on the intervening-test, $t(112) = .91$, $p = .37$. For RETRIEVAL SUCCESS, Levene's test of homogeneity of variances was non-significant, $p > .05$. Participants who had LOW RETRIEVAL SUCCESS on the intervening-test had significantly lower WMC ($M = 25.24$, $SD = 8.14$) than those who had HIGH RETRIEVAL SUCCESS ($M = 29.30$, $SD = 7.56$), $t(174) = -3.43$, $p < .01$.

**DISCUSSION**

**Summary of Results**

The first question this research sought to answer was whether FFQs offer an additive benefit to learning by leveraging the testing effect through feedback that is, itself, a test. Descriptives hinted that giving FFQs as feedback on the intervening-test would result in higher retention than not giving any feedback or only showing the correct answer. Results of a one-way ANOVA, however, showed that there was no statistically significant difference in retention-test scores between the three feedback groups. The most surprising point in this result is the lack of significant difference in retention-test scores between the group who saw no feedback and the two groups who received feedback. Schmidt and Bjork (1992) found that, in some cases, more feedback hindered retention when the feedback became part of the
task during practice but was removed during the retention test, as was the case in the present study.

Within the two groups who saw feedback, an independent-samples t-test showed that seeing more feedback during the intervening-test did not aid in retention. Indeed, participants who saw less feedback had significantly higher retention-test scores than those that saw more feedback on the intervening-test. This sample’s failure to benefit significantly from more feedback helps to explain the non-significant differences in retention-test scores between the three groups. In the cases where less feedback was associated with higher retention-test scores, participants who saw less feedback were more similar than dissimilar to the participants in the NO FEEDBACK group.

Viewed from the perspective of success instead of failure, participants who did well on the intervening-test saw less feedback than those who did poorly. Across all participants, an independent-samples t-test revealed that retention-test score was significantly higher among participants who had high retrieval success on the intervening-test than those who had low retrieval success. A one-way ANOVA showed that intervening-test score (i.e., retrieval success) was highest in the FFQ group. This makes logical sense given that only the participants in the FFQ group were given a chance to correct an erroneous initial answer.

Considering that intervening-test scores were significantly higher for participants who saw less feedback, the effect of the FFQ’s 2nd chance is more likely due to increasing the opportunity for successful retrieval and not to more exposure to feedback. Isolating the FFQ data, a Pearson’s product-moment correlation confirmed this conclusion. For example, successfully answering question A on the retention-test was related to successful retrieval of
question A’s answer on the intervening-test. This held true for the short-answer questions retrieved successfully by participants in the FFQ group, as shown in the last Pearson’s product-moment correlation.

The second question this research sought to answer was whether working memory capacity (WMC) moderated the strength or directionality of the relationship between type of feedback and intervening-test score. WMC did not significantly affect the relationship. Despite this non-significant moderation of intervening-test score between the three feedback groups, an independent samples t-test revealed that participants who had high retrieval success, regardless of feedback group, also had higher WMC than participants who had low retrieval success on the intervening-test. Another way to interpret this result is that participants who had significantly higher WMC also had significantly higher retrieval success.

**Implications**

Taken as a whole, these findings lead to a conclusion that FFQs offer an additive benefit to learning by leveraging the testing effect through feedback that is, itself a test. The implication of this conclusion to the retrieval effort theory, as it relates to feedback, lies in the success of the retrieval effort. Failed retrieval, regardless of level of retrieval effort, does not benefit learning. However, if feedback in the form of a question (e.g., FFQs) results in successful retrieval, then the amount of retrieval effort employed in answering the feedback question may yield more or less retention benefits. From a practical perspective, the use of FFQs have the potential to improve learning outcomes beyond what has been found in testing effect studies that have included feedback that is not, itself, a test.
Limitations

One major limitation of the present study was the use of fairly unsophisticated automated scoring. Mis-spellings such as “aqeous” for “aqueous” and “C02” for “CO2” were common causes of missed points and presentation of feedback, where applicable.

A possible limitation of the present study is that there may be some effect of motivation of participants because the payout was dependent on completing the task, not for performance. If this method were used in a real world setting, e.g., a classroom, then the results of the tests may be more meaningful to the participants and might inspire more effort. Speculatively, increased effort (and in some cases desperation) might enhance the effect of FFQ or suppress the beneficial effect of FFQs with a higher initial score.

In addition, the geology article contained difficult content that would be appropriate in an upper-level college geology course, and this prevented much prior knowledge from being useful. This was intentional for the purpose of preventing ceiling effects, which would have suppressed results. However, there may be differences in how participants respond to FFQs when the content is more familiar or easily accessible.

Continuing this line of thought, the present study only required participants to demonstrate declarative knowledge. However, FFQs might be equally or more effective for tests of procedural knowledge, such as math, physics, or programming. In these cases, the FFQs could be targeted to the discrete elements of the process with varying degrees of specificity.

Finally, this study did not examine the factor structure of the initial items or FFQs, and so there is no account of the possible families of items that might differentially
contribute to the final retention-test score. Such families of items might clump together due to similarities in content or item wording that only become obvious after performing an exploratory factor analysis. For example, if participants tended to fail all items relating to molecular dating even when given FFQs, then there may be cause to speculate that molecular dating might be too complex. Or it could be that FFQs that are verbatim to the article form one factor and items that are adjusted form another. Knowing possible factors could inform future research.

**Future Research**

If, as this study suggests, retention varies most powerfully by amount of retrieval success, then methods to increase retrieval success should be further investigated. One method of doing this might be to administer a series of FFQs with decreasing levels of difficulty, such that the leap between the final FFQ and the correct answer is very small. This premise might enable us to explore at what point retrieval effort once again becomes predictive of retention, when retrieval success is more or less guaranteed at the lowest level of retrieval effort.
REFERENCES


Appendix A: Mechanical Turk Informed Consent

Principal Investigator: Mary Luong
Faculty Sponsor: Dr. Christopher Mayhorn

INFORMED CONSENT FORM for RESEARCH
North Carolina State University

A Geology Learning Study

What are some general things you should know about research studies? You are being asked to participate in a research study. Your participation in this study is voluntary. You have the right to be a part of this study, to choose not to participate or to stop participating at any time. The purpose of research studies is to gain a better understanding of a certain topic or issue. You are not guaranteed any personal benefits from being in a study. Research studies also may pose risks to those that participate. In this consent form you will find specific details about the research in which you are being asked to participate. If you do not understand something in this form it is your right to ask the researcher for clarification or more information. You may print this consent form using your Web browsers print option. If at any time you have questions about your participation, do not hesitate to contact the researcher named below.

What is the purpose of this study? You are invited to participate in a research study. The goal of this study is to determine the quality of the educational content contained in a geology article.

What will happen if you take part in the study? If you agree to participate in this 2-part study, you will perform an initial task, be asked to answer a few questions about yourself (such as your gender, age), take a short test about geology, read a geology article, take a test about the information you read in the article, wait 1 week, then return and take a final test about the geology article. This study should take approximately 45 minutes of your time.

Risks: There are no risks or discomforts associated with this study.

Benefits: This research will be used to help us understand how to successfully design educational articles that help people learn.

Confidentiality: The information in the study records will be kept confidential to the full extent allowed by law. Data will be stored securely in a secured file. No reference will be made in oral or written reports which could link you to the study. You will NOT be asked to enter your name or any personally identifying information in any study materials so that no one can match your identity to the answers that you provide. Your computer IP address is not captured or stored.
**Compensation:** You will receive the compensation in two parts. The first payment of $0.50 will be awarded at the end of Part 1 of this study. The final payment of $1.00 will be paid, using Mechanical Turks bonus payment feature, when you return after the 1-week waiting period and complete Part 2 of this study. At the end of each part, a code will be provided that you must enter into Mechanical Turk to confirm you completed that part.

What if you have questions about this study? If you have questions at any time about the study or the procedures, you may contact the researcher, Mary Luong at mgluong@ncsu.edu, or Dr. Christopher Mayhorn at chris_mayhorn@ncsu.edu.

What if you have questions about your rights as a research participant? If you feel you have not been treated according to the descriptions in this form, or that your rights as a participant in research have been violated during the course of this project, you may contact Deb Paxton, Regulatory Compliance Administrator, Box 7514, NCSU Campus (919.515.4514).

**CONSENT TO PARTICIPATE:** By clicking on the I CONSENT option below, you are agreeing with the following statement: I understand that I have been asked to participate in a study. I have read and understand the above information. I know that I can print a copy of this form. I agree to participate in this study with the understanding that I may choose not to participate or to stop participating at any time without penalty or loss of benefits to which I am otherwise entitled.

NOTE: Participants were actually awarded a bonus of $1.50 for completing Part 2 due to feedback from the Mechanical Turk Workers about the low pay rate for the HIT.
Appendix B: Undergraduate Psychology Course Informed Consent

Principal Investigator: Mary Luong
Faculty Sponsor: Dr. Christopher Mayhorn

INFORMED CONSENT FORM for RESEARCH
North Carolina State University

A Geology Learning Study

What are some general things you should know about research studies? You are being asked to participate in a research study. Your participation in this study is voluntary. You have the right to be a part of this study, to choose not to participate or to stop participating at any time. The purpose of research studies is to gain a better understanding of a certain topic or issue. You are not guaranteed any personal benefits from being in a study. Research studies also may pose risks to those that participate. In this consent form you will find specific details about the research in which you are being asked to participate. If you do not understand something in this form it is your right to ask the researcher for clarification or more information. You may print this consent form using your Web browsers print option. If at any time you have questions about your participation, do not hesitate to contact the researcher named below.

What is the purpose of this study? You are invited to participate in a research study. The goal of this study is to determine the quality of the educational content contained in a geology article.

What will happen if you take part in the study? If you agree to participate in this 2-part study, the 1st PART will take about 40 minutes and include
- an initial task,
- answering a few questions about yourself (such as your gender, age),
- taking a short test about geology,
- reading a geology article,
- and taking a test about the information you read in the article.

For PART 2, you must
- wait 1 week,
- then return and take a final test about the geology article. Part 2 should take approximately 5 minutes of your time.

Risks: There are no risks or discomforts associated with this study.

Benefits: This research will be used to help us understand how to successfully design educational articles that help people learn.
Confidentiality: The information in the study records will be kept confidential to the full extent allowed by law. Data will be stored securely in a secured file. No reference will be made in oral or written reports which could link you to the study. You will NOT be asked to enter your name or any personally identifying information in any study materials so that no one can match your identity to the answers that you provide. Your computer IP address is not captured or stored.

Compensation: You will receive extra credit in your course but you MUST COMPLETE BOTH PARTS to get your extra credit.

What if you have questions about this study? If you have questions at any time about the study or the procedures, you may contact the researcher, Mary Luong at mgluong@ncsu.edu, or Dr. Christopher Mayhorn at chris_mayhorn@ncsu.edu.

What if you have questions about your rights as a research participant? If you feel you have not been treated according to the descriptions in this form, or that your rights as a participant in research have been violated during the course of this project, you may contact Deb Paxton, Regulatory Compliance Administrator, Box 7514, NCSU Campus (919.515.4514).

CONSENT TO PARTICIPATE: By clicking on the I CONSENT option below, you are agreeing with the following statement: I understand that I have been asked to participate in a study. I have read and understand the above information. I know that I can print a copy of this form. I agree to participate in this study with the understanding that I may choose not to participate or to stop participating at any time without penalty or loss of benefits to which I am otherwise entitled.
Appendix C: Demographics Survey

Please indicate your gender.
Male
Female
Prefer Not To Answer

Please indicate the highest level of education you have completed:
Some high school
High school
Some college
Undergraduate degree
Graduate degree

Please indicate your ethnic heritage.
African/African American
Asian/Asian American
Caucasian
Hispanic
Native American/Alaskan Native
Native Hawaiian/Pacific Islander
Other
Prefer Not To Answer

Is English your first language?
Yes
No

How confident are you that you could pass (70% or higher) a geology test?
Very Confident
Somewhat Confident
Not Confident

How much do you know about geology?
Somewhere between nothing and what I remember from high school/college
I have an undergraduate degree in geology
I have an advanced degree in geology
I study geology as a hobby
Appendix D: Devil’s Hole Fact Sheet

What is Devils Hole?

Devils Hole is a subaqueous cavern in south-central Nevada within a geographically detached unit of Death Valley National Park. The cavern is tectonic in origin and has developed in Cambrian carbonate rocks bordering the Ash Meadows oasis. The open fault zone comprising the cave extends to a depth of at least 130 meters below the water table, which is about 15 meters below land surface. The primary source of groundwater flowing through Devils Hole, and discharging from the major springs within the oasis, is precipitation on the Spring Mountains to the east of the cavern.

Why is Devils Hole of interest to paleoclimatologists?

The importance of Devils Hole to paleoclimatologists is twofold. Below the water table, the near-vertical walls of Devils Hole are coated with up to 40 centimeters of vein calcite that precipitated from groundwater moving through the cavern. The calcite can be accurately and precisely dated with radiometric methods, such that the depth-varying sequences of oxygen and carbon stable isotopes in the calcite provide a record of climatic variations spanning more than 500,000 years. Additionally, rippling subhorizontal cave deposits, called folia, record variations of the water table during the past 120,000 years.

The folia are porous deposits that owe their origin to outgassing, or release of gas, at the water table of CO2 from the slightly supersaturated groundwater flowing through Devils Hole. In Brown’s Room, a remote portion of Devils Hole cave, the folia were sampled up to about 9 meters above the modern water table. Uranium-series dating of these folia has enabled the development of a 120,000-year hydrograph of water-table fluctuation in the cavern, which indicates a prominent decline during the past 20,000 years.

What paleoclimate phenomena are recorded by the Devils Hole stable isotopic time series?

The Devils Hole oxygen-18 (δ18O) time series is primarily a proxy indicator of paleotemperatures. Unlike oxygen isotopes in deep-sea cores, it is not a record of past global ice accumulation in terrestrial systems. Rather, the time series appears to correspond, both in timing and relative magnitude, to variations in paleo-sea-surface temperature (SST) recorded in Pacific Ocean sediments off the west coast of North America, from Oregon to California, and as far south as the equatorial eastern Pacific. The Devils Hole δ18O record is also highly correlated with major variations in paleotemperatures recorded in the Vostok ice core from the East Antarctic Plateau. The Devils Hole carbon-13 (δ13C) time series is thought to reflect changes in global variations in the ratio of stable carbon isotopes of atmospheric carbon dioxide (CO2) and (or) changes in the density of vegetation in the groundwater-recharge areas tributary to Devils Hole.
What contributions has Devils Hole research made to the fields of paleoclimatology, paleohydrology, and geochemistry?

Publication of the half-million-year-long Devils Hole δ18O time series in 1992 initiated a decades-long discussion regarding the capability of the Milankovitch hypothesis (also referred to as orbital or astronomical theory) to predict the onset and duration of Pleistocene ice ages, a discussion that continues to this day. The hypothesis is named after Milutin Milankovitch, who mathematically theorized that variations in eccentricity, axial tilt, and precession of the Earth's orbit determined climatic patterns on Earth. The radiometrically dated Devils Hole time series indicates an atmospheric warming beginning more than 10,000 years earlier than the timing of the penultimate major deglaciation (also referred to as Termination II) as predicted on the basis of the Milankovitch hypothesis. This has been considered to be a direct challenge to the validity of orbital theory.

This challenge was dismissed by some on the grounds that if the earliest indications of warming were ignored, the Devils Hole records might be interpreted as supporting the Milankovitch hypothesis. But this view has been contradicted by others who claimed that early warming, though real, reflected only regional, not global, climate. All in all, the original challenge posed by the early warming recorded in the Devils Hole δ18O time series continues: it has been reinforced by a variety of studies, not only of interhemispheric paleo-SST warming, but also of sea-level high stands (a direct proxy for global ice volume conditions) that have been directly dated as occurring thousands of years prior to orbitally predicted global deglaciations.

Prior to publication of the Devils Hole δ18O time series, the duration of Pleistocene interglacial intervals was thought to be on the order of 12,000 years, as determined from orbitally tuned chronologies based on marine sediment records. However, an examination of the Devils Hole record published in 1997 revealed that the warmest portion alone of the past four interglaciations lasted on the order of 10,000 to 15,000 years, whereas the entire duration of the warm intervals (interglacials) persisted in excess of 20,000 years, a finding supported by both Antarctic ice core data and sea level data.

More recent work has corroborated these insights derived from Devils Hole, and it is widely acknowledged today that astronomical forcing alone cannot explain the duration, intensity, or diversity of the past four interglacials. More independently dated proxy records from both hemispheres would be required to achieve such an understanding. In addition, it is now widely acknowledged that variations in atmospheric greenhouse gases, notably CO2, have played a major role in the glacial-interglacial climatic shifts of the Pleistocene, though major questions persist regarding the origin of the “ice ages.”
Appendix E: Geology Questions

Q1: In terms of its location relative to water, what type of cavern is Devils Hole?

FFQ1: Try that again ... Extending to a depth of at least 130 meters below the water table, Devil’s Hole is classified as a ______ cavern.

A1: In terms of its location relative to water, what type of cavern is Devils Hole?
The correct answer is: subaqueous

Q2: What is the primary source of the groundwater found in Devils Hole?

FFQ2: Try that again ... The primary source of groundwater flowing through Devils Hole, and discharging from the major springs within the oasis, is ____ on the Spring Mountains.

A2: What is the primary source of the groundwater found in Devils Hole?
The correct answer is: precipitation

Q3: What is another name for the rippling cave deposits in Devils Hole?

FFQ3: Try that again ... Rippling subhorizontal cave deposits, called ____, record variations of the water table during the past 120,000 years.

A3: What is another name for the rippling cave deposits in Devils Hole?
The correct answer is: folia

Q4: What mineral in Devils Hole provides paleoclimatologists a record of ancient climate variations?

FFQ4: Try that again ... Depth-varying sequences of oxygen and carbon stable isotopes in the ____ provide a record of climatic variations spanning more than 500,000 years.

A4: What mineral in Devils Hole provides paleoclimatologists a record of ancient climate variations?
The correct answer is: calcite
Q5: The Devils Hole oxygen-18 (δ18O) time series appears to correspond to what type of variations?

FFQ5: Try that again ... The Devils Hole δ18O record is highly correlated with major variations in ____ recorded all the way from Pacific Ocean sediments to the Vostok ice core from the East Antarctic Plateau.

A5: The Devils Hole oxygen-18 (δ18O) time series appears to correspond to what type of variations?

The correct answer is: paleotemperatures (“paleotemp” was accepted for scoring purposes)

Q6: What stable isotopic time series reflects changes in the density of vegetation near Devils Hole?

FFQ6: Try that again ... The Devils Hole ____ time series is thought to reflect changes in global variations in the ratio of stable carbon isotopes of atmospheric carbon dioxide (CO2) and (or) changes in the density of vegetation in the groundwater-recharge areas.

A6: What stable isotopic time series reflects changes in the density of vegetation near Devils Hole?

The correct answer is: carbon-13 (“carbon 13” or “δ13C” were accepted)

Q7: What is another term for the Milankovitch hypothesis?

FFQ7: Try that again ... The Milankovitch hypothesis (also referred to as _____ or _____ theory) theorized that variations in eccentricity, axial tilt, and precession of the Earth's orbit determined climatic patterns on Earth.

NOTE: YOU ONLY NEED TO PROVIDE (1) One ALTERNATE TERM, not both.

A7: What is another term for the Milankovitch hypothesis?

The correct answer is: orbital or astronomical theory (“orbit” or “astronom” were accepted)

Q8: Prior to the publication of the Devils Hole oxygen-18 (δ18O) time series, how many years was the duration between interglacial intervals thought to be?

FFQ8: Try that again ... Prior to the publication of the Devils Hole oxygen-18 (δ18O) time series, the duration of Pleistocene interglacial intervals was thought to be on the order of ____ years, as determined from orbitally tuned chronologies based on marine sediment records.

A8: Prior to the publication of the Devils Hole oxygen-18 (δ18O) time series, how many years was the duration between interglacial intervals thought to be?

The correct answer is: 12,000
Q9: The subject of this test is Devils Hole. What is the subject of this test?

FFQ9: The subject of this test is Devils Hole. Type the following response below: Devils Hole.

A9: The subject of this test is Devils Hole. What is the subject of this test? **The correct answer is:** Devils Hole

*Note: Question 9 was an experimental check that the participants were paying attention and was not included in the intervening-test scores. Interestingly, about 5% of participants got this question wrong.

Q10: According to the Devils Hole record, entire interglacials lasted for how many years?

FFQ10: Try that again ... An examination of the Devils Hole record published in 1997 revealed that the entire duration of the warm intervals (interglacials) persisted in excess of _____ years, a finding supported by both Antarctic ice core data and sea level data.

A10: According to the Devils Hole record, entire interglacials lasted for how many years? **The correct answer is:** 20,000

Q11: What atmospheric gas has played a role in glacial-interglacial climatic shifts?

FFQ11: Try that again ... It is now widely acknowledged that variations in atmospheric greenhouse gases, notably ____, have played a major role in the glacial-interglacial climatic shifts of the Pleistocene.

A11: What atmospheric gas has played a role in glacial-interglacial climatic shifts? **The correct answer is:** CO₂ ("CO2" or "carbon dioxide" were accepted)

Q12: Although some suggest that the Devils Hole time series supports the Milankovitch hypothesis, critics of that view attribute early warming to what kind of climate change?

FFQ12: Try that again ... If the earliest indications of warming were ignored, the Devils Hole records might be interpreted as supporting the Milankovitch hypothesis. But this view has been contradicted by others who claimed that early warming, though real, reflected only _____, not global, climate.

A12: Although some suggest that the Devils Hole time series supports the Milankovitch hypothesis, critics of that view attribute early warming to what kind of climate change? **The correct answer is:** regional ("region" was accepted)

Q13: Folia are formed by CO₂ in a process called what?
FFQ13: **Try that again ...** The folia are porous deposits that owe their origin to ____, or release of gas, at the water table of CO2 from the slightly supersaturated groundwater flowing through Devils Hole.

A13: Folia are formed by CO2 in a process called what?  
**The correct answer is:** outgassing (“outgas” was accepted)

Q14: The folia in Devils Hole indicate that the water table was once at least how many meters high?

FFQ14: **Try that again ...** In Brown’s Room, a remote portion of Devils Hole cave, the folia were sampled up to about ____ meters above the modern water table.

A14: The folia in Devils Hole indicate that the water table was once at least how many meters high?  
**The correct answer is:** 9

Q15: Devils Hole developed in what kind of rocks?

FFQ15: **Try that again ...** Devils Hole is tectonic in origin and has developed in ____ ____ rocks bordering the Ash Meadows oasis.

A15: Devils Hole developed in what kind of rocks?  
**The correct answer is:** Cambrian carbonate

Q16 Dating the Devils Hole folia has produced what type of diagram of water-table fluctuation in the cavern?

FFQ116: **Try that again ...** Uranium-series dating of Devils Hole folia has enabled the development of a 120,000-year ____ of water-table fluctuation in the cavern, which indicates a prominent decline.

A16: Dating the Devils Hole folia has produced what type of diagram of water-table fluctuation in the cavern?  
**The correct answer is:** hydrograph

*Note: This question was not included in the intervening-test scores because in the CORRECT ANSWER condition the answer was not provided due to error on my part when setting up this question in Qualtrics.*

47
Q17 Where is Devils Hole located?

FFQ17: Try that again ... Devils Hole is in south-central _____ within a geographically detached unit of Death Valley National Park.

A17: Where is Devils Hole located?
The correct answer is: Nevada
Appendix F: Debriefing Document

Title of Study: Facilitative Feedback Questions: Refining the Retrieval Effort Theory of the Testing Effect

Principal Investigator: Mary Luong  Faculty Sponsor: Dr. Christopher Mayhorn

This study is concerned with the retrieval effort theory of the testing effect. The study was designed to assess the power of facilitative feedback questions to improve long-term retention of the geology article. Facilitative feedback provides elaboration that allows learners to correct themselves.

Everyone read the same geology article and was asked the same questions. Some people got no feedback, others were given the correct answer when they gave an incorrect answer, and still others were given facilitative feedback questions when they gave an incorrect answer.

If you would like to receive a report of this research when it is completed (or a summary of the findings), please contact Mary Luong at mgluong@ncsu.edu.

If you have concerns about your rights as a participant in this experiment, please contact Deb Paxton, Regulatory Compliance Administrator, NCSU Campus Box 7514 (919-515-4514).

Please do not disclose research procedures and hypotheses to anyone who might participate in this study between now and the end of the data collection (December, 2013) as this could affect the results of the study. Thank you for your participation!