THE ROLE OF MATERIAL COMPLEXITY IN RETRIEVAL PRACTICE EFFECTS

by

Joseph P. Bedwell

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Psychological Sciences

West Lafayette, Indiana

May 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Jeffrey D. Karpicke, Chair
Department of Psychological Sciences
Dr. Thomas S. Redick
Department of Psychological Sciences
Dr. James S. Nairne
Department of Psychological Sciences

Approved by:
Dr. David Rollock
Head of the Graduate Program

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Bedwell, Joseph P. MS
Institution: Purdue University
Degree Received: May 2018
Title: The Role of Material Complexity in Retrieval Practice Effects
Major Professor: Jeffrey D. Karpicke

Recently, van Gog and Sweller (2015) asserted that the advantages of testing diminish (or even disappear altogether) as the complexity of learning materials increases. To elucidate their claim, they used the term "element interactivity" as a proxy for material complexity. While material low in element interactivity can be thought of as a series of isolated facts, material that is high in element interactivity contains strongly related ideas such that the learning of any particular idea is contingent on understanding other components of the material. The current set of experiments systematically manipulated element interactivity in order to evaluate the validity of van Gog and Sweller's contention.

Experiment 1 manipulated element interactivity by scrambling the order of sentences within educational texts. Specifically, students studied two educational texts that were either presented intact or with their sentences scrambled. For one of the passages, students engaged in retrieval practice following study, and for the other, they completed a distractor task following study. Subjects' memory for the passage content was assessed a week later when they were asked to answer a series of questions about the texts and freely recall the information they studied the previous week. Although it may seem counterintuitive, the logic put forth by van Gog and Sweller (2015) would argue that for the scrambled passages (which were lower in element interactivity) testing effects would be present, and for the passages presented with their sentences intact (higher

element interactivity), testing effects should be absent. Contrary to this prediction, doing retrieval practice led to enhanced performance on all sections of the final test regardless of whether the texts were intact or scrambled.

Experiment 2 manipulated element interactivity by altering subjects' contextual prior knowledge. Participants studied an ambiguously worded passage in the presence or absence of a topic word that provided a relational schema to guide their interpretation. Because the topic word served to create relations among the ideas presented in the text, van Gog and Sweller's hypothesis would assert that retrieval practice effects should be absent when the topic word is visible and present when it is not visible. As in Experiment 1, subjects either did retrieval practice or completed a distractor task after studying the passage and took a final test roughly one week later. Results showed a benefit of retrieval practice for both the topic absent and the topic present groups.

Experiment 3 was largely identical to Experiment 2 in that it manipulated subjects' contextual prior knowledge and followed a similar procedure. However, in this case, participants studied word lists that conformed to ad hoc categories in the presence or absence of the category names. In terms of predictions, we speculated that giving subjects' access to the category names would increase element interactivity, thereby implying that van Gog and Sweller (2015) would argue that testing effects should be absent when category names are present. In stark contrast to this assertion, findings from this experiment indicated a benefit of engaging in retrieval practice regardless of whether the category names were present or not. Nevertheless, there were no differences in performance between the category names present and category names absent groups.

Thus, across three experiments, no evidence was found to support the contention that retrieval practice effects are absent when using complex materials.

INTRODUCTION

The past decade has seen an explosion of research concerning the topic of retrieval-based learning (Karpicke, 2017). Indeed, this is not without cause because the testing effect, the finding that information that is repeatedly retrieved from memory is better remembered than information that is repeatedly studied, has myriad educational implications and appears to be quite robust. For instance, the benefits of testing (more commonly referred to as retrieval practice) seem to transcend age, such that its effectiveness is evident in preschool aged children (Fritz, Morris, Nolan, & Singleton, 2007) and elderly adult populations (Bishara & Jacoby, 2008; Logan & Balota, 2008). Additionally, retrieval practice effects generalize to many educational contexts and disciplines, including biology (Carpenter et al. 2016), foreign language (Karpicke & Roediger, 2008), psychology (Batsell, Perry, Hanley, & Hotstetter, 2017), general science (Karpicke & Blunt, 2011), and statistics (Lyle & Crawford, 2011). Although the plethora of aforementioned studies might lead one to conclude that the testing effect is ubiquitous, recently a debate has emerged concerning whether the complexity of the learning materials serves as a boundary condition for the phenomenon.

In general, the effects of retrieval practice persist across a range of materials. However, as one might infer, the bulk of the research has been conducted using materials that are easily verbalized. To give a few examples, testing effects have been found in studies that used educational texts (Blunt & Karpicke, 2014; McDaniel, Howard & Einstein, 2009; Wissman, Rawson, & Pyc, 2011), key-term definitions (Ariel & Karpicke, 2018; Grimaldi & Karpicke, 2014), word-pair associates (Carpenter 2009; Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke & Bauernschmidt, 2011), and word

lists (Carpenter & DeLosh, 2006; Whiffen & Karpicke, 2017). Moreover, the benefits of retrieval practice extend to materials that are difficult to articulate. Kang (2010) asked subjects to learn a series of Chinese characters by repeatedly studying them or by engaging in retrieval practice through mental visualization. On a final test, the advantage of the retrieval practice condition over the restudy condition was evident. Finally, spatial learning tasks also seem to receive a benefit from testing. Kelly, Carpenter, and Sjolund (2015) had subjects learn a route that was to be used to navigate through a series of connected rooms in a virtual building. After being presented with the correct route once, subjects either restudied it or were asked to recall the route from memory. The subjects who had their memory tested exhibited superior performance on a final test, provided feedback was given after they selected a doorway.

Despite the wealth of research investigating the effects of retrieval practice with different material types, the relationship between testing effects and material complexity has been infrequently evaluated. Recently, van Gog and Sweller (2015) asserted that the testing effect decreases (or even disappears entirely) when the complexity of the learning materials increases. To elucidate their claim, they used the term “element interactivity” as a proxy for material complexity. Material low in element interactivity can be thought of as a series of isolated facts where any single idea can be learned without reference to the other ideas in the material. By contrast, material that is high in element interactivity contains strongly related ideas such that the learning of any particular idea is contingent upon understanding other components of the material.

As evidence for their contention that the testing effect is diminished when using materials that are high in element interactivity, van Gog and Sweller (2015) cited prior

research involving worked examples. Worked examples describe or illustrate the procedure for solving a particular problem. Often, these problems consist of multiple steps, so the primary function of a worked example is to delineate the order in which the sequence of steps should be accomplished. Additionally, because each step builds on the previous one, worked examples are classified as high element interactivity (van Gog et al. 2015; Leahy et al. 2015). Therefore, to examine the relationship between retrieval practice effects and high element interactivity materials, van Gog et al. (2015) had students learn to solve electrical circuit problems using worked examples. Specifically, students were randomly assigned to one of three conditions. The first group studied four different worked examples, the second group alternated between studying a worked example and solving an isomorphic problem (the procedure for solving the problem was the same, but there were different surface features), and the third group alternated between studying a worked example and then solving the problem they previously studied. One week later, participants took a final test consisting of isomorphic circuit problems, and no differences between the groups were found. Because this result was consistent with earlier worked example studies that used a similar procedure, the authors concluded that the testing effect is absent for high element interactivity materials (van Gog & Kester, 2012; van Gog et al. 2015).

Although further examination into the relationship between worked examples and testing effects is certainly warranted, it is worth noting that van Gog and Sweller's claim was met with some opposition (Karpicke & Aue, 2015; Rawson 2015). Most notably, critics took issue with the fact that none of the worked example experiments manipulated element interactivity. Instead, the authors presented a series of studies in which element

interactivity was supposedly high and testing effects were absent. A more compelling set of data, then, would be one that manipulates element interactivity, holds other factors constant, and shows that retrieval practice effects are present when element interactivity is low and reduced or absent when element interactivity is high.

In an effort to satisfy these criteria, Hanham, Leahy, and Sweller (2017) had elementary and middle school students learn to write poems by studying worked examples. Students were assigned to one of two groups. The first group studied two example poems and then produced their own poem. The second group studied one example poem, then filled in blanks with the rules for composing the poem (retrieval practice), and finally wrote their own poem. Crucially, the rules for constructing a poem either tapped low or high element interactivity information and depending on the extent to which a particular rule was followed on the final test, the type of element interactivity knowledge the rule was associated with was considered learned. For instance, a sample low element interactivity rule was “there must be six lines in the poem,” and if the student’s final poem contained six lines, they were considered to have successfully completed a low element interactivity part of the task. An example high element interactivity rule was “the first letter in the first five lines must form a word.” Results indicated that a greater proportion of the low element interactivity information was satisfied when students did retrieval practice, but a greater proportion of the high element interactivity information was satisfied when students studied two examples. However, this pattern of results only occurred on an immediate test. For a later experiment that used a delayed test, there was a numerical advantage for the subjects that engaged in retrieval practice in terms of satisfying both low and high element interactivity criteria.

Additionally, this experiment was underpowered with fewer than 15 subjects per cell, which likely influenced whether a significant difference was detected.

Regardless of the extent to which van Gog and Sweller's assertion is supported in experiments using worked examples, further investigation into the relationship between element interactivity and retrieval practice effects is justified. Specifically, any conclusions drawn will have substantial educational implications. While students need to learn both low and high element interactivity knowledge in order to succeed in the classroom, information that is high in element interactivity is more commonly thought of as the goal to strive for in that it implies students will be able to see connections between ideas. Furthermore, less complex material often serves as a prerequisite for understanding material of greater complexity, and as such, educators should be made cognizant of whether there is a point when practicing retrieval becomes an ineffective strategy for learning new content. Conversely, if material complexity does not function as a boundary condition for the testing effect, educators should be informed of this additional facet of its ubiquity and remain steadfast in their use of retrieval practice within the classroom.

Material Complexity as Text Structure

While material complexity can be manipulated in a variety of ways, one technique that several studies have used is altering the structure of a text. For instance, in their second experiment, Karpicke and Blunt (2011) sought to determine whether retrieval practice served as a more effective learning strategy than creating a concept map when studying two texts that differed in terms of their structure. Specifically, they presented subjects with an enumeration passage (a text that listed a series of facts and concepts) and a sequential passage (a text that described a sequence of interrelated events). Arguably,

the sequential passage would be higher in element interactivity than the enumeration passage, such that one might expect differences in performance depending on the passage structure. However, the results of the experiment revealed a large benefit of retrieval practice over concept mapping regardless of whether the passage was classified as sequential or enumeration. A similar finding was demonstrated by Blunt and Karpicke (2014) when equivalent retrieval practice effects were found irrespective of whether subjects were presented with a sequential or an enumeration passage.

Although van Gog and Sweller (2015) did not personally conduct any studies that manipulated element interactivity, they reported that scrambling the order of sentences in a passage reduces it. de Jonge, Tabbers, and Rikers (2015) investigated this idea by asking subjects to study a passage about black holes that was either intact (sentences presented in a coherent order) or scrambled (sentences presented in an incoherent order). To confirm that this manipulation was effective, the degree of coherence within each text was assessed through Latent Semantic Analysis, and this revealed that the two texts contained significantly different coherence levels. Therefore, two separate experiments were conducted. The first presented subjects with the intact text in a sentence-by-sentence format. Following 15 minutes of initial study, subjects either restudied the passage for 15 minutes or were tested on the passage content for 15 minutes. In the test condition, students were re-presented with each of the passage's sentences individually and asked to fill in blanks that indicated missing words. Regardless of their condition, subjects returned to the lab one week later and took a final fill-in-the-blank test. Results of this experiment indicated that performance on the final test was the same in both the study and test conditions. The second experiment followed the exact same procedure

except that students were presented with the incoherent version of the text. In this experiment, a benefit of testing was revealed such that on the delayed assessment, students exhibited less forgetting in the test condition than in the restudy condition.

A similar experiment was conducted by Chan (2009) where students were asked to read educational texts. Crucially, the sentences within each paragraph were either presented intact or randomly-ordered, a feature which Chan referred to as high and low integration, respectively. Students studied each passage for 16 minutes, but for one of the texts they also engaged in retrieval practice of the content by answering a series of short-answer questions that required them to relate multiple concepts within the material. Contrary to the findings of de Jonge et al. (2015), Chan observed robust benefits of retrieval practice on delayed tests one day after the initial learning phase for both the low and high integration conditions.

Given the conflicting results presented above, it is necessary to consider how the structure of a text could influence retrieval practice effects. Drawing upon the "elaborative retrieval" (Carpenter, 2009) and "mediator strengthening" (Kornell, Klein, & Rawson, 2015) hypotheses, van Gog and Sweller (2015) argued that the benefit of retrieval lies in its ability to establish relations between information elements and provide an organizational structure to the material. Therefore, they hold that an advantage of testing would be evident in texts with lower element interactivity (i.e., those where the sentences have been scrambled) due to the increased relational processing elicited by retrieval. Conversely, texts higher in element interactivity (which contain an inherent organizational structure) receive no added benefit from practicing retrieval because the relational processing it affords is redundant with the structure of the passage.

The idea that complex materials render the testing effect impotent as a consequence of redundant processing is interesting when one considers the literature concerning material-appropriate processing. A material-appropriate processing framework argues that the greatest enhancements to recall performance occur when the processing elicited by the learning activity is complimentary to the type of processing afforded by the text (McDaniel & Einstein, 1989, 2005; McDaniel & Butler, 2010). Specifically, it identifies two types of processing: relational and item-specific. While relational processing emphasizes similarity and highlights the importance of making connections between multiple ideas, item-specific processing underscores distinctiveness and the unique features of individual items (Grimaldi, Poston, & Karpicke, 2015). Within the framework, both types of processing are needed to produce maximum recall, and if the types of processing invited by the text and learning activity are the same, then lower levels of recall will be observed.

The decision to invoke material-appropriate processing as an explanation for why increased material complexity reduces the benefit of retrieval practice seems imprudent, given the emphasis van Gog and Sweller (2015) placed on the de Jonge et al. (2015) article. To recap, in that experiment, the learning activity involved filling in individual words within each sentence, and the text was either presented intact or with the sentences scrambled. Therefore, if one assumes that the scrambled text affords less relational processing than the coherent text and that the activity of filling in individual words relies on item-specific processing, it is reasonable to expect that a testing effect would be evident when the passage was intact and that it would be absent when the text was scrambled. The opposite pattern of results was found. Hence, it seems that if material

complexity does serve as a boundary condition for retrieval practice effects, it will be necessary to consider alternative explanations.

Although the multitude of studies that found a testing effect using coherent educational texts cast doubt on the claim that retrieval practice does not benefit highly relational material, there is good evidence to suggest that testing does enhance organizational processing. When asked to recall uncategorized word lists repeatedly, it has been demonstrated that subjects recall according to a subjective order, continually placing the same words next to each other on successive test trials (Tulving, 1962, 1966). A similar finding has been revealed with categorized word lists. Bregman and Wiener (1970) had subjects study the lists and then recall them on three consecutive test trials. While the proportion of words recalled was the same across test trials, category clustering increased across the successive tests—a finding that has been consistently replicated (Congleton & Rajaram, 2012; Zaromb & Roediger, 2010). Notably, because this trend toward organization was found in studies that used categorized and uncategorized word lists, it gives credence to the idea that retrieval practice benefits can be found despite a preexisting organizational structure.

If the advantages of testing are independent of whether the material is presented in a coherent manner, can this fit into currently proposed mechanisms for retrieval practice? One possibility is the episodic context account (Karpicke, Lehman, & Aue, 2014). This explanation holds that when individuals encode material, they also encode information about the temporal context in which the information is presented. When that piece of material is retrieved at a later point in time, individuals attempt to reinstate the prior temporal context. If they are successful, the context representation associated with

the item or material is updated, such that it incorporates contextual features from the time it was originally studied and the time it was recalled. This allows individuals to restrict their search set and have several effective contextual retrieval cues to help them access the content. Crucially, this account would suggest that because temporal context information is independent of any preexisting organizational scheme within a passage, the benefits of retrieval would persist regardless of the structure of a text.

Material Complexity as Contextual Prior Knowledge

Another factor that can influence the complexity of a set of material is whether an individual has access to the requisite contextual information. This idea was illustrated quite clearly in a series of experiments conducted by Bransford and Johnson (1972) where subjects were asked to recall information that made little sense unless appropriate schematic knowledge was given. In the first experiment, high school students listened to an audio recording of a text, but depending on the condition to which the student was assigned, they were either presented with a picture that gave the text meaning before they heard the recording, after they heard the recording, or not at all. Importantly, the passage that the subject heard was written to follow standard rules of English language construction, except that the sentences themselves were vague and had ambiguous interpretations. Furthermore, the passage did not attempt to describe the picture, but rather the picture served as a contextual base for where the events described in the passage could occur. Regardless of whether a picture was presented, all subjects attempted to recall the passage and rate their ease of comprehension following a two-minute delay. Results found that there was no difference in recall and comprehension ratings between the individuals who were not exposed to the picture and those who were

shown the picture after the recording of the passage was played. However, when contextual information was given prior to hearing the passage, the proportion of idea units recalled more than doubled and comprehension ratings significantly improved.

In their subsequent experiments, Bransford and Johnson (1972) used a similar procedure that required subjects to listen to an audio recording of a text and then attempt to recall the content. Critically, the method used to provide subjects with contextual information was altered, such that in these cases instead of being presented with a picture, students were given a one-word topic that described the passage (e.g., “laundry”). Again, because the statements from the passage were quite nebulous, the authors felt that the presence of the topic would assist the subjects in developing a schema to organize the content. The pattern of results mirrored the previous experiment; when the topic was presented before the passage was heard, recall performance was more than twice as large as it was when the topic was presented after the audio recording had finished or when no topic was given at all. While neither of these experiments could be considered an examination of the testing effect (due to the lack of restudy controls), they do suggest that initial recall performance can be aided through the provision of contextual information.

Aside from providing subjects with a passage topic, several other methods of imparting contextual information have been examined. For example, giving subjects background information to read (Barnett, 1984; Rawson & Kintsch, 2002), granting them access to informational outlines (Mannes & Kintsch, 1987), and other advance organizers (Corkill, 1992; Dunlosky, Rawson, & Hacker, 2002) have all shown to improve subjects’ recall performance. Interestingly, van Gog and Sweller (2015) would likely classify manipulations such as these as ones that increase the element interactivity or complexity

of the passage due to their ability to provide individuals with a relational schema that they can use to link the ideas in the text. However, these studies are similar to Bransford and Johnson (1972) in that none of them directly investigated the testing effect, but because they did lead to higher levels of recall, it suggests that there is potential for a benefit of repeated retrieval to exist in such a scenario.

Unlike investigations of material complexity that involve text structures, examinations of material complexity that involve prior contextual knowledge can also be evaluated through recall of word lists. Bower (1970) found that preventing subjects from establishing consistent organizational structures (by altering their perceptual groupings) hindered free recall performance. Specifically, compared to subjects who were given a consistent organizational scheme, individuals who experienced words presented in differing perceptual groupings learned a much smaller proportion of items across several multi-trial recall attempts. Couple this finding with the aforementioned evidence that people tend to cluster words recalled into categories, and it is clear that having an organizational structure plays an essential role in the recall of word lists.

One way to assess how the presence or absence of contextual information affects the recall of word lists is to use lists that conform to an ad hoc category. Ad hoc categories are composed of words that can be linked in an atypical fashion (e.g., things made of metal, things that are green), such that subjects generally perceive them as being unrelated (Hunt & Einstein, 1981; McDaniel, Moore, & Whiteman, 1998). Therefore, when subjects attempt to study and recall the list, it is unlikely that they will be aware of its organizational structure unless they are made cognizant of the categories beforehand. For example, Little, Lewandowsky, and Heit (2006) conducted a study in which

participants were asked to recall words that fit into particular ad hoc categories. In this case, the experimenters manipulated whether the subjects were told of the category names before or after their recall attempts. They found that when participants were told of the ad hoc categories after their recall attempts, on subsequent trials their categorization strategy changed to reflect this newfound knowledge. Similar to providing subjects with the topic of the passage they are studying, a manipulation such as this would likely serve to increase the perceived relatedness and element interactivity of the content.

Distinct from the small base of literature that investigated the relationship between the testing effect and the structure of a text, there appears to be a dearth of studies that evaluate the relationship between retrieval practice and contextual prior knowledge. With that said, both topics fall within the realm of material complexity, and as such the mechanisms of retrieval that were applied to text structures will also be discussed here. If complex materials do serve as a boundary condition for the effects of retrieval, it is reasonable to consider a material-appropriate processing approach as van Gog and Sweller (2015) suggest. By its nature, free recall is a learning activity that evokes relational processing, and it is quite possible that if contextual information is provided that helps subjects to see the connections within a set of material, the two instances of relational processing will be redundant and diminish the testing effect. Conversely, if contextual information functions to increase the testing effect, one could argue that the contextual background knowledge works to create a more potent retrieval cue that subjects can reinstate during recall. Finally, if testing serves to improve performance regardless of whether contextual information is provided, it lends credence

to the episodic context account in that the temporal context information is independent of other types of context information that could be potentially reinstated.

Introduction to the Experiments

In order to further elucidate the relationship between material complexity and retrieval practice effects, three experiments were conducted. The first investigated material complexity by manipulating text structure, whereas the latter two manipulated material complexity by altering subjects' contextual prior knowledge. More specifically, the first experiment aimed to clarify the role that scrambling passage sentences has on testing effects, the second looked at the effects of retrieval when subjects were presented with an ambiguous passage in the presence or absence of a topic word to guide their interpretation, and the third examined how knowledge of word lists' ad hoc categories influence the effects of testing. In all cases, subjects either engaged in retrieval practice by repeatedly studying and recalling the content or participated in a control condition that required them to perform distractor tasks during the time when retrieval would be taking place. In addition, all experiments incorporated a one-week delay, during which participants' knowledge of the material was assessed a final time, thereby serving as a means to evaluate the enduring effects of retrieval. Regardless of their outcome, these studies seek to provide definitive evidence as to whether the benefits of testing interact with or are independent of material complexity. Moreover, because the experiments pertain to the potential ubiquity of the testing effect, there are substantial educational implications, as well as opportunities to inform our theoretical explanations of retrieval practice.

EXPERIMENT 1

Experiment 1 was conducted to investigate how material complexity influences retrieval practice effects when it is defined as the degree of coherence within a text. In this instance, the coherence of the text was manipulated by presenting subjects with passages that were either intact or had their sentences scrambled. In order to determine whether this manipulation was effective, the degree of referential cohesion within each version of the text was measured using Coh-Metrix, a program that provides multilevel analysis of text characteristics (Graesser, McNamara, & Kulikowich, 2011). In that context, referential cohesion is defined as the extent to which ideas within a text overlap and are connected across adjacent sentences, a definition that underscores its capability to serve as an index of element interactivity within a passage.

Assuming the effectiveness of our text coherence manipulation, the potential for a testing effect was enabled by having subjects read two educational texts. For one of the passages, students engaged in repeated retrieval, alternating between studying the text and attempting to freely recall its content (two cycles). For the other passage, subjects completed a distractor task after each study phase, such that the total time spent engaging in retrieval and completing the distractor task was equivalent. To ascertain whether there was a benefit of retrieval, subjects returned to the lab one week later and answered a series of questions about the passages and attempted to recall each text once. Given that van Gog and Sweller (2015) asserted that the testing effect is diminished as the element interactivity of the learning materials increases and that scrambling sentences is an effective method for reducing element interactivity, an evaluation of the benefit of retrieval for both the scrambled and intact texts was of critical importance.

If complex materials do function as a boundary condition for the testing effect, we would expect no benefit of retrieval to be observed for the intact texts and an advantage of retrieval to be found when the texts were scrambled (because they are lower in element interactivity). Conversely, if retrieval practice effects are only evident when the texts are presented intact, we can conclude that making material less cohesive (and less complex according to the element interactivity hypothesis) disrupts the gains elicited through retrieval. Finally, if an advantage of retrieval persists regardless of whether the materials are high or low in terms of their element interactivity, it offers evidence that material complexity is orthogonal to the testing effect.

Method

Subjects

Sixty Purdue University undergraduate students participated in this study in exchange for course credit. All subjects were fluent in written and spoken English, and the mean age of participation was 19.8.

Materials

Two brief texts, “Tropisms” and “Homeostasis,” were selected from Cook and Mayer (1988). Both passages were identified as having a “generalization” structure, meaning they were written in such a way that inferential thinking was promoted. Specifically, each text contained a central idea, and the sentences within the passage served to either explain that idea with illustrations or extend the idea with key details. Two versions of each passage (one with the sentences intact and another with the sentences scrambled) were created and run through Coh-Metrix in order to measure their degree of referential cohesion. Referential cohesion refers to overlap in content words

among adjacent sentences, and this is determined by calculating the average number of sentences where a given element (noun, argument, stem, and anaphor) overlaps with the previous sentence (Graesser, McNamara, & Kulikowich, 2011). For the tropisms passage, Coh-Metrix identified the original text as having a referential cohesion z-score of 1.11 and the scrambled version a z-score of 0.64, indicating that element interactivity was reduced.[1] The two versions of the text had word counts of 260 and 262 respectively, and the differences in word count emerged because in the scrambled version of the text anaphoric references were replaced with their corresponding nouns (e.g. "they" was replaced with "bean plants"). For the homeostasis passage, the referential cohesion z-scores were 0.03 and -0.68 for the intact and scrambled versions of the text, respectively, indicating that once again the manipulation of element interactivity was successful. In this case, the intact passage contained 265 words and the scrambled passage consisted of 284 words. The two versions of the "Tropisms" and "Homeostasis" passages can be found in Appendices C and B respectively.

Design

Each student studied two educational texts. The structure of the texts (intact vs. scrambled) was manipulated between-subjects. The two learning activities the students engaged in (retrieval practice vs. study and complete a distractor task) were manipulated within-subject, such that a different activity was completed for each text. The order in which the texts were presented was consistent across subjects, and the order of the two learning activities was counterbalanced. Thirty subjects were randomly assigned to each of the between-subjects conditions.

---

[1] *Z*-scores were calculated by comparing the passages to a corpus of 37,520 texts provided by Touchstone Applied Science Associates.

Procedure

Experiment 1 consisted of two sessions. In both sessions, subjects were tested in small groups of up to seven individuals at a time. Each student sat at a computer, and all elements of the task as well as their instructions were presented on-screen. Regardless of whether subjects were assigned to view intact or scrambled versions of the text, Session 1 of the experiment began with the student studying the "Tropisms" passage for 4 minutes. During study sessions, the title of the text was centered and bolded at the top of the screen with the contents of the text below it. Students were instructed to read the text for the entire 4-minute period and were told that the experiment would advance automatically after that time had elapsed.

Following the initial study period, subjects either engaged in a retrieval attempt or completed a distractor task for 8 minutes. During instances of retrieval, subjects were shown a screen with the title of the text they were supposed to recall bolded and centered at the top. Beneath this was a response box where students were instructed to type as much of the content from the text as they could remember. Subjects were encouraged to continue typing for the entire 8-minute period and were told the experiment would advance automatically once that time had passed. During instances where students were asked to complete a distractor task, they played a video game for the 8-minute period. Again, once this interval was complete, the experiment advanced automatically. Following their initial attempt at retrieval or their first instance of completing the distractor task, students were instructed to study the passage again for 4 minutes. After this, they completed whichever activity they did earlier (retrieval or distractor) a second time for 8 minutes.

The second passage presented to subjects was entitled “Homeostasis.” The cycle that students completed for the first passage was essentially repeated here in that subjects studied the passage for 4 minutes, completed an activity for 8 minutes, restudied the passage for 4 minutes and completed the activity again for 8 minutes. The only difference was that this time around subjects completed the activity they were not exposed to for the first passage. After their final 8-minute interval had finished, subjects were dismissed from Session 1 and thanked for their participation.

Session 2 occurred exactly 1 week after Session 1 and was the same for all participants. The session began with subjects answering a series of 12 short answer questions pertaining to the “Tropisms” passage. Six of the questions were identified as verbatim questions in that they quizzed students over information that was explicitly stated in the text. The other six questions were denoted inferential and as such required that students make connections between multiple ideas and apply the passage content. Short answer questions were presented individually on the screen, and subjects were required to spend at least 15 seconds trying to answer the question. After that time elapsed, a “continue” button appeared that students could use to advance to the next question. However, it should be noted that this portion of the experiment was entirely self-paced and students were encouraged to spend as much time answering the question as they felt was necessary. Upon completion of the short answer questions, subjects were prompted to recall the “Tropisms” passage for 8 minutes. The format for this recall trial was identical to the recall trials in Session 1. After 8 minutes had passed, students repeated the procedure this time answering 12 questions about the “Homeostasis”

passage and spending 8 minutes attempting to recall its contents. Once this final recall session was complete, subjects were debriefed and thanked for their participation.

Results

Scoring

All responses to the short answer questions were scored by two independent raters. Raters were instructed to score responses as either correct (1 point), partially correct (0.5 points) or incorrect (0 points). To help ensure grading consistency and delineate distinctions among response categories, each rater was given a rubric that contained the correct response to each question as well as reasons for assigning partial credit. The two raters gave the same score to 93% of the responses, and in the event that there was a discrepancy, a third rater cast the deciding vote. In order to score the free recall responses, each text was broken down into a series of 30 idea units. Again, two independent raters scored all responses, assigning full, partial, or no credit depending on the extent to which any given idea unit was present in the student's answer. No credit was awarded when an idea unit was completely absent from a student's response, partial credit was given if the subject partially referenced an idea unit without stating it explicitly, and full credit was assigned if the idea unit was fully present in the response. The independent raters were in agreement on 90% of the recall responses, and in the instances that a disagreement occurred, a third rater was called in to resolve the conflict.

Initial Recall Performance

A preliminary analysis indicated that there was a difference in initial recall performance between passages, such that recall was higher on the "Tropisms" text than on the "Homeostasis" text. However, this did not interact with any other factors aside

from final recall performance, so the results have been collapsed across texts. Table 1 shows the proportion of idea units recalled in each period for both the intact and scrambled texts. Collapsed across text structure, the proportion of ideas recalled increased from Period 1 to Period 2 (.30 vs. .47), $t(59) = 8.75$, $p < .001$, $d = 1.13$, 95% confidence interval (CI) [0.74, 1.51]. In addition, students recalled more idea units from the intact text than from the scrambled text. This pattern occurred in Period 1 (.36 vs. .24), $t(58) = 2.48$, $p = .016$, $d = 0.64$, 95% CI [0.12, 1.16] and in Period 2 (.54 vs. .39), $t(58) = 2.53$, $p = .016$, $d = 0.65$, 95% CI [0.13, 1.17].

Short Answer Performance

Figures 1 and 2 show performance on the final short answer test that occurred one week after the initial session for verbatim and inference questions respectively. For each passage, subjects answered 6 verbatim questions and 6 inference questions; therefore, results are broken down by question type. For verbatim questions, a 2 (learning condition: study and complete a distractor task vs. study and do retrieval practice) by 2 (text structure: intact vs. scrambled) mixed factorial ANOVA revealed a main effect of learning activity, $F(1, 58) = 23.37$, $p < .001$, $\eta_p^2 = .29$ and a main effect of text structure, $F(1, 58) = 6.00$, $p = .017$, $\eta_p^2 = .09$. Importantly, there was no learning activity by text structure interaction, $F(1, 58) = 0.231$, $p = .631$. Specifically, for the intact texts, performance on the verbatim short answer questions was higher for students who studied and did retrieval practice than for students who studied and completed a distractor task, (.75 vs. .60), $t(28) = 3.06$, $p = .003$, $d = 0.80$, 95% confidence interval (CI) [0.27, 1.32]. The pattern persisted for the scrambled texts; performance on the verbatim short answer questions was higher for students who studied and did retrieval practice than for

studentswho studied and completed a distractor task, (.61 vs. .43) $t(28) = 3.73$, $p < .001$, d =0.96, 95% CI [0.42, 1.49]. In sum, while performance on verbatim questions was lower when the texts were scrambled, retrieval practice effects were found regardless of whether the passages were intact or scrambled.

The results were quite similar for the inferential questions. A 2 by 2 mixed factorial ANOVA revealed a main effect of learning activity, $F(1, 58) = 32.16$, $p < .001$, $\eta_p^2 = .36$, a main effect of text structure, $F(1, 58) = 6.51$, $p = .013$, $\eta_p^2 = .10$, and no learning activity by text structure interaction, $F(1, 58) = 2.58$, $p = .114$. Specifically, for the intact texts, performance on the inferential questions was higher for students who studied and did retrieval practice than for students who studied and completed a distractor task, (.69 vs. 49), $t(28) = 5.10$, $p < .001$, $d = 1.32$, 95% CI [0.75, 1.87]. For the scrambled passages, performance on the inferential questions was also higher for students who studied the passage and did retrieval practice than for students who studied and completed a distractor task, (.50 vs. .39), $t(28) = 2.85$, $p = .006$, $d = 0.74$, 95% CI [0.21, 1.26]. Again, this indicates that while subjects had better performance on the inference questions if they studied the intact texts, there was a benefit of retrieval practice regardless of whether the texts were intact or scrambled.

Final Recall Performance[2]

A preliminary analysis indicated that there was a difference in final recall performance between passages, such that recall was higher on the “Tropisms” text than on the “Homeostasis” text. However, this did not interact with any other factors aside

[2]It is difficult to interpret performance on the final recall test because a confound exists in that students always answered a series of short answer questions about the text prior to completing the final recall.

from initial recall performance, so the results have been collapsed across texts. Table 2 shows the proportion of idea units recalled for each text as a function of learning activity and text structure. As in the earlier analyses, a 2 (learning activity) by 2 (text structure) mixed factorial ANOVA was used to determine the presence of main effects and interactions. This analysis indicated that there was a main effect of learning activity, $F(1, 58) = 46.97$, $p < .001$, $\eta_p^2 = .71$, a main effect of text structure $F(1, 58) = 4.86$, $p = .032$, $\eta_p^2 = .08$, and no interaction, $F(1, 58) = 1.70$, $p = .198$.

Specifically, for the intact texts, performance on the final recall was higher for students who studied and did retrieval practice than for students who studied and completed a distractor task, (.48 vs. 30), $t(28) = 5.78$, $p < .001$, $d = 1.49$, 95% CI [0.91, 2.06]. The pattern persisted for the scrambled texts; performance on the final test was higher for students who studied and did retrieval practice than for students who studied and completed a distractor task, (.33 vs. .21), $t(28) = 3.93$, $p < .001$, $d = 1.01$, 95% CI [0.47, 1.55]. In a manner consistent with the short answer data, this analysis demonstrates that final free recall performance was aided by studying intact texts rather than scrambled texts. Moreover, despite the particular text structure subjects were exposed to, their performance improved when they used retrieval practice instead of studying and completing a distractor task.

Discussion

The purpose of Experiment 1 was to determine whether altering the structure of a text (making it more or less complex) would influence the presence and potency of a testing effect. Specifically, students studied two educational texts that were either presented intact or with their sentences scrambled. For one of the passages, students

alternated between studying the text and free recall attempts (2 cycles), and for the other, they completed a distractor task instead of doing free recall. Following a one-week delay, subjects returned to the lab to complete a final free recall of each text and answer a series of questions about the passages. According to van Gog and Sweller's element interactivity hypothesis, the benefit of testing would be diminished when the texts were presented intact due to their increased complexity. Contrary to that assertion, the results of Experiment 1 found an advantage for the texts where students used retrieval practice regardless of whether the passages were intact or scrambled. Furthermore, even though performance on the final test was lower when texts were scrambled, this did not serve to reduce or eliminate the testing effect, thereby casting doubt on the claim that retrieval does not enhance the learning of complex materials.

EXPERIMENT 2

The results of Experiment 1 provided strong evidence that element interactivity, when defined as the degree of coherence or structure within a text, had little bearing on the efficacy of retrieval practice. In an attempt to extend these findings, Experiment 2 sought to address a related question: To what degree is the presence and potency of the testing effect influenced by material complexity when it is operationalized as an individual's contextual prior knowledge? To that end, Experiment 2 manipulated contextual prior knowledge by presenting subjects with a passage in either the presence or absence of a topic word that provided a relational schema. Specifically, subjects were asked to study a passage that, in very vague terms, described the process of doing laundry. In this case, half of the participants were exposed to the topic word, "laundry" when viewing the passage and half were not. Of crucial importance to the current design is that all subjects were presented with the exact same passage; the only factor that changed was whether they had a relational schema to interpret the passage. Because van Gog and Sweller (2015) operationalized material complexity as element interactivity and classified highly interactive material as more complex, this study provides a means of manipulating material complexity without altering the material itself.

Consistent with Experiment 1, Experiment 2 investigated the presence of a retrieval practice benefit by either prompting subjects to recall the material after study or to complete a distractor task following study. Then, on a delayed test approximately one week after the initial learning phase, subjects attempted to freely recall the passage content. Because this manipulation was coupled with the presence or absence of the topic word, this experiment sought to provide insight into whether the testing effect is

influenced by contextual prior knowledge. Importantly, if material complexity does serve as a boundary condition for retrieval practice effects (as the element interactivity hypothesis asserts) we would expect a learning activity by topic word interaction such that the benefit of retrieval is smaller when subjects are presented with the topic word than when they are not given that relational framework. Alternatively, if the topic word seems to exaggerate the advantage of retrieval, it is plausible to reason that its presence serves to create more potent contextual retrieval cues that an individual can reinstate. Finally, a result similar to Experiment 1 would give cause for an analogous interpretation: the effect of material complexity (when defined as contextual prior knowledge) is orthogonal to the testing effect.

Method

Subjects and Design

Experiment 2 consisted of a 2 (topic word: absent vs. present) by 2 (learning activity: retrieve vs. study and distractor) between-subjects design. Therefore, four distinct conditions were produced. G*Power was used to determine a sample size large enough to detect difference in the size of the testing effect up to effect sizes of $d = 0.60$ (which is common in studies of the testing effect) with a power of 0.80 and an alpha level of .05 (Faul, Erdfelder, Lang, & Buchner, 2007). This analysis indicated that 45 subjects were needed for each of the four conditions.

Two hundred and twenty-one subjects were recruited through an online Human Intelligence Task (HIT) posted on TurkPrime (Litman, Robinson, & Aberbock, 2017). Eligible subjects were restricted to individuals who lived in the United States, had a HIT acceptance rate of 95% or greater, and had completed at least 1000 HITs. Of the 221

individuals who completed Session 1 of the experiment, 187 returned for Session 2. Six of those subjects were removed because they self-identified as having cheated at some point during the experiment, and one additional participant was removed because their first language was not English. Therefore, the final sample consisted of 180 subjects, with 45 randomly assigned to each of the four between-subjects conditions. Demographically speaking, the sample was comprised of 87 females and 94 males, and 153 individuals identified their race as white. In terms of age, participants ranged from 18 to 68 (M = 35.4, SD = 9.4). For Experiment 2, subjects completed two online sessions, the second occurring 6-8 days after the subject completed Session 1. The majority of subjects (71%) completed Session 2 exactly six days after Session 1. In terms of compensation, each subject received $3.50 ($1.50 for Session 1 and $2.00 for Session 2). The duration of Session 1 was approximately 15 minutes, so subjects were paid at a rate of 10 cents per minute. Session 2 lasted approximately 6 minutes, but subjects were paid extra to incentivize them to return for the session.

Material

The passage entitled “Laundry” was taken from Bransford and Johnson (1972). The text contains 162 words. If subjects were assigned to a condition which necessitated their exposure to the topic word, the word “Laundry” was centered and bolded above the passage. The exact passage that the participants studied and attempted to recall can be found in Appendix E.

Procedure

Experiment 2 consisted of two sessions. Session 1 was initiated after subjects accepted the HIT on the TurkPrime website. Prior to accepting the HIT, subjects were

informed of the two-part nature of the experiment and notified that they were expected to return and complete Session 2 within 6-8 days of finishing Session 1. Provided the HIT was accepted, subjects read and electronically signed a consent form and filled out demographic information. From this point onward, the procedure for the experiment differed slightly depending on the condition to which the subject was assigned. However, regardless of their condition, each subject was presented with detailed instructions prior to every phase of the experiment that outlined what was expected of them for the upcoming task. All participants began the experiment with a study phase in which they were tasked with reading a text for 2 minutes. If subjects were assigned to a topic present condition, the word “Laundry” was written in bold and centered above the text. For the topic absent conditions, no such title was provided. Each participant was instructed to study the passage for 2 minutes and told that the experiment would advance automatically after that time had elapsed.

Following the initial study period, subjects either engaged in a retrieval attempt or completed a distractor task for 4 minutes. During instances of retrieval, subjects were shown a screen with the word “Recall” centered and in bold at the top. Beneath this was a response box in which participants were instructed to type as much of the content from the text that they could remember. Subjects were encouraged to continue typing for the entire 4-minute period and were told the experiment would advance automatically once that time had passed. During instances where subjects were asked to complete a distractor task, their digit span was tested for 4 minutes. The digit span task presented participants with a sequence of numbers ranging from 4-9 digits in length at a rate of 1 digit per second. After all of the digits in a sequence had been shown, a calculator appeared on

screen and subjects were asked to input the digits they saw in the exact order they were presented. Once the participant was satisfied with the numbers they had selected, they pressed a continue button to advance to the next number sequence. In addition, to encourage engagement with the distractor task, participants were told to complete as many number sequences with the highest possible degree of accuracy within the 4-minute period. Once this interval was complete, the experiment advanced automatically.

After their initial attempt at retrieval or their first instance of completing the distractor task, participants were instructed to study the passage again for 2 minutes. After this, they completed whichever activity they did earlier (retrieval or distractor) a second time for 4 minutes. Session 1 was concluded after this 4-minute interval. Subjects were given a brief survey asking whether they cheated (they were assured they would receive payment regardless of their response) or if they had additional comments. They were then thanked for their participation and reminded to complete Session 2 in 6-8 days.

Six days after they completed Session 1, subjects received an email from TurkPrime informing them that the second part of the experiment was available. If participants did not respond to this initial invitation, they were sent reminder emails every 6 hours until the 6-8 day window was over. Session 2 was identical for all participants. They were prompted to recall the passage they studied 6-8 days ago for 4 minutes, and the topic word, “Laundry” was not shown at any point during the second session. After this was complete, subjects completed the same survey they filled out at the end of Session 1, presented with debriefing information, and thanked for their participation.

Results

Scoring

In a manner similar to Experiment 1, the text was broken down into a series of 14 idea units and responses were scored by two independent raters. However, unlike the previous experiment each idea unit was scored as either present or absent (1 point vs. 0 points); no partial credit was given. The independent raters were in agreement on 96% of the free recall responses, and in the instances where there was a discrepancy, the two raters scores were averaged together resulting in 0.5 points being awarded for that particular idea unit.

Initial Recall Performance

Table 3 shows the proportion of idea units recalled in each period for both the topic present and topic absent groups. Collapsed across those groups, the proportion of idea units recalled increased from Period 1 to Period 2 $t(89) = 9.51$, $p < .001$, $d = 0.71$, 95% CI [0.50, 0.92]. In addition, while subjects recalled a numerically greater proportion of the idea units when the topic was present, this difference was not significant at Period 1 (.51 vs. .44), $t(88) = 1.41$, $p = .162$ and at Period 2 (.66 vs. .60), $t(88) = 1.21$, $p = .230$.

Final Recall Performance

Figure 3 shows the proportion of idea units recalled during the final recall session as a function of learning activity and topic word presence. A 2 (learning activity) by 2 (topic word) between-subjects ANOVA revealed a main effect of learning activity, $F(1, 176) = 37.20$, $p < .001$, $\eta^2 = .174$, a main effect of topic word $F(1, 176) = 5.75$, $p = .018$, $\eta^2 = .032$, and no interaction, $F(1, 176) = 1.47$, $p = .228$.

Specifically, when the topic word was present, performance on the final recall was higher for participants who studied and did retrieval practice than for participants who studied and completed a distractor task, (.40 vs. 15), $t(88) = 5.19$, $p < .001$, $d = 1.09$ 95% CI [0.65, 1.53]. The pattern persisted when the topic word was absent; performance on the final recall was higher for individuals who studied and did retrieval practice than for individuals who studied and completed a distractor task, (.28 vs. .11), $t(88) = 3.48$, $p = .001$, $d = 0.73$ 95% CI [0.30, 1.16]. In a manner consistent with Experiment 1, this analysis demonstrates that final recall performance improved when subjects engaged in retrieval practice regardless of whether the topic word was present or not. Furthermore, the magnitude of this effect was greater for the topic present group, which is significant given that particular condition was thought to be higher in element interactivity/material complexity.

Laundry-Related Intrusions

To potentially offer some insight as to whether the presence of the topic word affected subjects' final recall performance as a function of the particular learning activity they engaged in, the number of laundry-related intrusions within each group were tallied. Laundry-related intrusions occurred when the subject included, in their free recall response, the specifics of a step involved in the laundry process that was not identified in the passage (e.g. adding detergent). Therefore, these intrusions represent a failure of the topic word in assisting a participant with restricting their search set. Consequently, if retrieval practice serves to help individuals restrict their search set, we would expect a lower number of laundry-related intrusions in the condition where subjects engaged in retrieval practice than the study and distractor condition. When subjects engaged in

retrieval practice and had the topic present during study, 2 of the 45 participants produced laundry-related intrusions on their final recall response. Compared to the 9 out of 45 subjects who produced laundry-related intrusions on the final recall after studying the passage with the topic present and engaging in a distractor task, a difference between the groups is evident, $\chi(1) = 5.08$, $p = .024$. Notably, in both conditions the maximum number of laundry-related intrusions produced by a single subject was 1. Hence, it appears that retrieval practice plays a role in reducing intrusions.

Discussion

The purpose of Experiment 2 was to determine whether altering a subject's contextual prior knowledge (making the material more or less complex) would influence the presence and potency of a testing effect. Specifically, subjects studied an ambiguously worded passage that detailed the steps of doing laundry in the presence or absence of a topic word that provided a relational schema. Learning activity was manipulated between-subjects; participants either alternated between studying the text and free recall attempts (2 cycles), or they completed a distractor task instead of doing free recall. Following a 6-8 day delay, subjects completed a final free recall of the text they studied roughly one week prior. According to van Gog and Sweller's element interactivity hypothesis, the benefit of testing would be diminished when the text was presented in the presence of the topic word (due to its ability to inter-relate the passage content). Contrary to that assertion, the results of Experiment 2 found an advantage when subjects used retrieval practice regardless of whether topic word was present or absent. Furthermore, the magnitude of the testing effect was greater when the topic word was present, thereby casting doubt on the claim that retrieval does not enhance the learning of

complex materials. Finally, there appears to be an association between learning activity and number of intrusions, such that retrieval practice effectually limited the number of intrusions brought about by the presence of the topic word.

EXPERIMENT 3

Experiment 3 sought to extend and potentially generalize the findings from Experiment 2 to a new type of material. To clarify, for this experiment material complexity was still operationalized as individuals' contextual prior knowledge, but in this case, word lists served as the to-be-learned information instead of text passages. To accomplish this, word lists that conformed to particular ad hoc categories were presented to the subjects, and the availability of contextual prior knowledge was determined by whether the participants had access to the category names during study trials. Significantly, this manipulation ensured that all subjects were presented with the exact same material and that the only difference was whether the participants had a relational schema with which they could interpret the material. Moreover, as in the previous experiments, the effect of retrieval was ascertained by having subjects either alternate between studying and recalling the words for 2 cycles or studying the words and completing a distractor task for an amount of time equivalent to that of the recall task. Finally, to assess the enduring impacts of the manipulations, a final test was given roughly one week after the initial session.

An additional factor to consider that was not pertinent in the first two experiments is whether asking subjects to retrieve words in the presence of ad hoc category names would lead to generation effects (Jacoby, 1978). Put simply, if the strategy differs at the time of retrieval such that subjects who have access to the category names are merely generating words that fulfill the criteria, then there is reason to expect smaller retrieval practice effects. Hypothetically, this could result because individuals who are just generating words are not engaging in context reinstatement to the same extent as their

counterparts who are not exposed to the category names. With that said, the word lists chosen for this experiment were selected because they contain relatively few words and the categories themselves are broad enough that the likelihood of randomly generating one is quite small.

According to the element interactivity hypothesis, presenting subjects with the ad hoc category names should increase the relational nature/complexity of the material. Therefore, in terms of testing effect and material complexity interactions, the element interactivity hypothesis would claim that the benefit of testing will be smaller when subjects have access to the category names during study (as a result of higher element interactivity). Conversely, if performance is higher when category names are available to subjects, it lends support to the assertion that the category names make it easier for subjects to reinstate a prior episodic context. Lastly, if an advantage of retrieval practice persists regardless of whether category names are present, we can conclude that there is little relationship between the testing effect and contextual prior knowledge.

Method

Subjects and Design

Experiment 3 consisted of a 2 (ad hoc category names: absent vs. present) by 2 (learning activity: retrieve vs. study and distractor) between-subjects design. Therefore, four distinct conditions were produced. Because the design of this experiment is similar to Experiment 2, the *a priori* power analysis used for that experiment is also relevant here. Consequently, we sought to have 45 subjects randomly assigned to each of the between-subjects conditions.

Two hundred and thirty-six subjects were recruited through an online HIT posted on TurkPrime. Eligible subjects were restricted to individuals who lived in the United States, had a HIT acceptance rate of 95% or greater, and had completed at least 1000 HITs. Of the 236 individuals who completed Session 1 of the experiment, 199 returned for Session 2. Eleven of those subjects were removed because they self-identified as having cheated at some point during the experiment, and eight additional participants was removed because their first language was not English. Therefore, the final sample consisted of 180 subjects, with 45 randomly assigned to each of the four between-subjects conditions. Demographically speaking, the sample was comprised of 78 females and 102 males, and 165 individuals identified their race as white. In terms of age, participants ranged from 21 to 71 (M = 34.9, SD = 13.1). For Experiment 3, subjects completed two online sessions, the second occurring 6-8 days after the subject completed Session 1. The majority of subjects (68%) completed Session 2 exactly six days after Session 1. In terms of compensation, each subject received $1.80 ($0.80 for Session 1 and $1.00 for Session 2). The duration of Session 1 was approximately 8 minutes, so subjects were paid at a rate of 10 cents per minute. Session 2 lasted approximately 3 minutes, but subjects were paid extra to incentivize them to return for the session.

Materials

An 18-word ad hoc categorized word list was used. The list contained three categories (a thing that makes noise, a thing that is green, a thing made of wood), with six words per category. The words were selected from the updated Battig and Montague (1969) norms (Van Overschelde, Rawson, & Dunlosky, 2004), and the words chosen for this experiment can be found in Appendix F.

Procedure

Experiment 3 consisted of two sessions. Session 1 was initiated after subjects accepted the HIT on the TurkPrime website. Prior to accepting the HIT, subjects were informed of the two-part nature of the experiment and notified that they were expected to return and complete Session 2 within 6-8 days of finishing Session 1. Provided the HIT was accepted, subjects read and electronically signed a consent form and filled out demographic information. From this point onward, the procedure for the experiment differed slightly depending on the condition to which the subject was assigned. However, regardless of their condition, each subject was presented with detailed instructions prior to every phase of the experiment that outlined what was expected of them for the upcoming task. All participants began the experiment with a study phase in which the 18 words were presented individually in random order. Specifically, each word was presented on-screen for 3 seconds with a 1 second inter-stimulus interval. If subjects were assigned to a condition where category names were present, a category name (the category name the word conformed to) was written in bold and centered above each individual stimulus. In these cases, the sentence, "Above each word you will see a category that the word conforms to written in bold" appeared in the instructions, but aside from that insertion, the instructions did not differ between groups. For the category names absent conditions, the category names were not provided. Each participant was instructed to study the words and not write anything down.

Following the initial study period, subjects either engaged in a retrieval attempt or completed a distractor task for 90 seconds. During instances of retrieval, subjects were shown a screen with the word "Recall" centered and in bold at the top. Beneath this was a

response box. Subjects were instructed to recall (in any order) as many words from the list that they were able to remember by typing them into the response box and pressing the “Enter” key after each entry. Additionally, participants were discouraged from guessing, and the program prevented them from submitting a word multiple times. Furthermore, all words added to the response box remained on-screen for the duration of the recall trial and subjects were unable to remove prior entries. All participants were encouraged to continue trying to recall words for the entire 90 second period and told the experiment would advance automatically once that time had elapsed.

During instances where subjects were asked to complete a distractor task, their digit span was tested for 90 seconds. The digit span task presented participants with a sequence of numbers ranging from 4-9 digits in length at a rate of 1 digit per second. After all of the digits in a sequence had been shown, a calculator appeared on screen and subjects were asked to input the digits they saw in the exact order they were presented. Once the participant was satisfied with the numbers they had selected, they pressed a continue button to advance to the next number sequence. In addition, to encourage engagement with the distractor task, participants were told to complete as many number sequences with the highest possible degree of accuracy within the 90 second period. Once this interval was complete, the experiment advanced automatically.

After their initial attempt at retrieval or their first instance of completing the distractor task, participants were instructed to studied each of the words again at a rate of 3 seconds/word. After this, they completed whichever activity they did earlier (retrieval or distractor) a second time for 90 seconds. Session 1 was concluded after this 90-second interval. Subjects were given a brief survey asking whether they cheated (they were

assured they would receive payment regardless of their response) or if they had additional comments. They were then thanked for their participation and reminded to complete Session 2 in 6-8 days.

Six days after they completed Session 1, subjects received an email from TurkPrime informing them that the second part of the experiment was available. If participants did not respond to this initial invitation, they were sent reminder emails every 6 hours until the 6-8 day window was over. Session 2 was identical for all participants. They were prompted to recall the words they studied 6-8 days ago for 90 seconds. The format for this recall trial was identical to the recall trials described earlier. After this was complete, subjects completed the same survey they filled out at the end of Session 1, were presented with debriefing information, and thanked for their participation.

Results

Scoring

All responses were scored automatically by a scoring algorithm that assigned 1 point if the word recalled was an exact match to one of the words on the list (differences in capitalization did not affect the score) or 0 points if a word recalled did not match one of the words on the list. To correct for spelling and pluralization errors, a single rater went back through the scored responses and assigned full credit to any item that differed from the correct spelling by three letters or fewer.

Initial Recall Performance

Table 4 shows the proportion of words correctly recalled in each period for both the category names present and category names absent groups. Collapsed across those groups, the proportion of words recalled increased from Period 1 to Period 2 $t(89) =$

12.45, $p < .001$, $d = 0.93$, 95% CI [0.71, 1.15]. In addition, there were not differences between the two groups in the number of words correctly recalled at Period 1 (.53 vs. .51), $t(88) = 0.36$, $p = .723$ and at Period 2 (.71 vs. .72), $t(88) = 0.28$, $p = .783$.

Final Recall Performance

Figure 4 shows the proportion of words correctly recalled during the final recall session as a function of learning activity and the presence of category names. A 2 (learning activity) by 2 (category names) between-subjects ANOVA revealed a main effect of learning activity, $F(1, 176) = 21.00$, $p < .001$, $\eta^2 = .107$, no main effect of category names $F(1, 176) = 0.21$, $p = .649$, and no interaction, $F(1, 176) = 0.00$, $p = .984$.

Specifically, when category names were present, performance on the final recall was higher for subjects who studied and did retrieval practice than for subjects who studied and completed a distractor task, (.26 vs. 11), $t(88) = 3.23$, $p = .001$, $d = 0.68$ 95% CI [0.25, 1.10]. The pattern persisted when the category names were absent; performance on the final was higher for individuals who studied and did retrieval practice than for individuals who studied and completed a distractor task, (.27 vs. .13), $t(88) = 3.25$, $p = .001$, $d = 0.69$ 95% CI [0.26, 1.11]. In a manner consistent with Experiments 1 and 2, this analysis demonstrates that final recall performance improved when subjects engaged in retrieval practice regardless of whether the category names were present or not.

Category Name Intrusions

Experiment 3 provided the opportunity to investigate whether the presence of category names influenced the prevalence of intrusions on the final test. Specifically, in the two conditions where subjects were exposed to category names during study, the number of times each subject recalled one of the category names on the final test (e.g.,

green) were tallied. Overall, the maximum number of category name intrusions produced by a single subject was 2. Furthermore, when subjects engaged in retrieval practice after studying with the category names, 9 intrusions were produced. This is compared to the 4 category name intrusions that occurred after subjects studied the words with category names present and completed a distractor task. A chi square analysis indicated that this difference between the two groups was not significant, $\chi(1) = 2.25$, $p = .134$. Thus, unlike Experiment 2, the number of intrusions subjects produced was not meaningfully influenced by the learning activity they completed.

Discussion

The purpose of Experiment 3 was to determine whether altering subject's contextual prior knowledge (making the material more or less complex) would influence the presence and potency of a testing effect. Specifically, subjects studied a series of words that conformed to particular ad hoc categories in the presence or absence of category names that provided a relational schema. Learning activity was manipulated between-subjects; participants either alternated between studying the text and free recall attempts (2 cycles), or they completed a distractor task instead of doing free recall. Following a 6-8 day delay, subjects attempted to freely recall the words they studied roughly one week prior. According to van Gog and Sweller's element interactivity hypothesis, the benefit of testing would be diminished when the text was presented in the presence of the topic word (due to its ability to inter-relate the content). Contrary to that assertion, the results of Experiment 3 found an advantage when subjects used retrieval practice regardless of whether category names were present or absent. However, there was no effect of category names on the final test, nor did there appear to be an association

between the number of intrusions produced on the final test and the particular learning activity the subject engaged in. Hence, altering whether category names were present or absent may not have been an effective manipulation of element interactivity.

GENERAL DISCUSSION

The purpose of these experiments was to test the assertion put forth by van Gog and Sweller (2015) that the testing effect is diminished when the complexity of learning materials increases. To ensure this claim was adequately evaluated, the term element interactivity, van Gog and Sweller's definition for material complexity, was adopted and systematically manipulated in three experiments. Material low in element interactivity can most easily be thought of as discrete facts where any single component of the material can be learned without reference to the other components of the material. Conversely, material high in element interactivity consists of inter-related ideas where the learning of any particular idea is contingent on understanding other pieces of information within the material. To clarify, element interactivity was treated as analogous to material complexity with high element interactivity representing high complexity and low element interactivity representing low complexity. Across three experiments, no evidence was found to support the contention that the testing effect is reduced with high element interactivity materials.

Experiment 1 manipulated element interactivity by scrambling the order of sentences within educational texts. This manipulation was chosen because it was identified by van Gog and Sweller (2015) as a technique for lowering element interactivity. Hence, students studied two educational passages that were either presented with their sentences intact or their sentences scrambled. Additionally, learning activity was manipulated within-subject such that for one of the passages the students engaged in retrieval practice following study, and for the other, they completed a distractor task following study. Subjects' memory for the passage content was assessed a week later

when they were asked to answer a series of questions about the texts and freely recall the information they studied the previous week. Although it may seem counterintuitive, the logic put forth by van Gog and Sweller (2015) would argue that for the scrambled passages (which were lower in element interactivity) testing effects would be present, and for the passages presented with their sentences intact (higher element interactivity), testing effects should be absent. Contrary to this prediction, doing retrieval practice led to enhanced performance on all sections of the final test regardless of whether the texts were intact or scrambled. Furthermore, performance was overall lower for the scrambled texts, but this did not preclude the presence of a testing effect.

The findings of Experiment 1 were closely mirrored by Experiment 2, except, in the case of the latter, element interactivity was manipulated by altering subjects' contextual prior knowledge. Specifically, participants studied an ambiguously worded passage in the presence or absence of a topic word that provided a relational schema to guide their interpretation. This manipulation was selected because of its ability to alter element interactivity without changing the to-be-learned content. Moreover, we reasoned that giving subjects the topic word would increase the element interactivity of the passage due to its ability to inter-relate the information. Therefore, van Gog and Sweller's hypothesis would assert that retrieval practice effects should be absent when the topic word is visible and present when it is not visible. Notably, Experiment 2 was similar to Experiment 1 in that participants either did retrieval practice after studying the passage or completed a distractor task after studying the passage. Additionally, a final test, in which the subjects were asked to freely recall the text, was given roughly one week following initial study. Even though performance was worse when the topic word was absent,

results showed a benefit of retrieval practice for both the topic absent and the topic present groups. Furthermore, the magnitude of the testing effect was larger when the topic word was present (the condition supposedly higher in element interactivity), which casts serious doubt on the claim that the testing effect is absent for complex materials.

Experiment 3 was largely identical to Experiment 2 in that it manipulated subjects' contextual prior knowledge and used a similar procedure. However, in this case, participants studied word lists that conformed to ad hoc categories in the presence or absence of the category names. Again, this type of manipulation was chosen because it altered the element interactivity of the material without changing the to-be-learned information. In terms of predictions, we speculated that giving subjects' access to the category names would increase element interactivity; hence, van Gog and Sweller (2015) would argue that testing effects should be absent when category names are present. In stark contrast to this assertion, findings from this experiment indicated a benefit of engaging in retrieval practice regardless of whether the category names were present or not. Nevertheless, there were no differences in performance between the category names present and category names absent groups. While this has little implication for the efficacy of the testing effect, one could argue that because of this, Experiment 3 was an ineffective manipulation of element interactivity. One possible reason for this outcome could be that the words from the word lists were presented individually, thereby making it more difficult for subjects to visualize the words that belonged to a particular category. A potential solution, then, could be to present all the words that conform to specific category simultaneously and see if that increases the strength of the manipulation.

Given that the benefit of retrieval practice seems to be independent of material complexity, it is necessary to consider how this finding can inform our theoretical understanding of the testing effect. In particular, these results appear consistent with the episodic context account of retrieval practice effects (Karpicke, Lehman, & Aue, 2014). This explanation holds that when individuals encode material, they also encode information about the temporal context in which the information is presented. When that piece of material is retrieved at a later point in time, individuals attempt to reinstate the prior temporal context. If they are successful, the context representation associated with the item or material is updated, such that it incorporates contextual features from the time it was originally studied and the time it was recalled. This allows individuals to restrict their search set and have several effective contextual retrieval cues to help them access the content. Importantly, this account would suggest that because temporal context information is independent of any of the contextual features of the material, one would not expect the efficacy of retrieval practice to be impacted by material complexity.

While this series of experiments was not designed to evaluate the episodic context account, it is perhaps possible to argue that the intrusion data from Experiment 2 offers additional support for the theory. Briefly, Experiment 2 required subjects to study a text about doing laundry, and an intrusion occurred if the presence of the topic word (laundry) caused subjects to describe a part of doing laundry that was not detailed in the passage. Findings from Experiment 2 indicated an association between the number of intrusions a subject produced and the learning activity they engaged in. Specifically, fewer intrusions occurred if subjects did retrieval practice than if they completed a distractor task following study. Notably, this finding illustrates that by doing retrieval practice subjects

were able to restrict their search set above and beyond any restrictions brought about by the presence of the topic word. Moreover, because the episodic context account emphasizes the role of search set restriction in retrieval practice effects, the viability of this theory should not be overlooked.

In sum, the results of three experiments offer evidence that the benefit of retrieval practice is independent of the complexity of learning materials. This finding has myriad educational implications in that it suggests to educators that self-testing is an effective strategy within the classroom regardless of the difficulty of the to-be-learned materials. Additionally, although these experiments provide a compelling look at the material complexity-retrieval practice relationship, they are by no means an exhaustive examination of the topic. Future research, then, should strive to build upon this work and investigate the effects of testing on different facets of material complexity.

LIST OF REFERENCES

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*(1), 43-56.

Barnett J. E. (1984) Facilitating retention through instruction about text structure. *Journal of Reading Behavior*, *16*, 1-13.

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology, 44* (1), 18-23.

Battig, W. F., & Montague, E. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs, 80*, 1-46.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic. Bulletin & Review, 15*(1), 52-57.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106,* 849-858.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology, 1*, 18-46.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717-726.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 9*, 689-698.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition, 34*(2), 268-276.

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D., (2016). A classroom study on the relationship between student achievement and retrieval enhanced learning. *Educational Psychology Review, 28*(2), 353-375.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153-170.

Congleton, A. R., & Rajaram, S. (2012). The origin of the interaction between learning history and delay in the testing effect: The roles of processing and retrieval organization. *Memory & Cognition, 40*, 528-539.

Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology, 80*, 448-456

Corkill, A. J. (1992). Advance organizers: Facilitators of recall. *Educational Psychology Review, 4*, 33-67.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review, 27*(2), 305-315.

Dunlosky, J., Rawson, K. A., & Hacker, D. J. (2002). Metacomprehension of science texts: Investigating the levels-of-disruption hypothesis. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *Psychology of science text comprehension* (pp. 255-279). Mahwah, NJ: Erlbaum.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology, 60*(7), 991-1004.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology, 106*, 58-68.

Grimaldi, P. J., Poston, L., & Karpicke, J. D. (2015). How does creating a concept map affect item-specific encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1049-1061.

Hanham, J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology, 31*(3), 265-280.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*, 497-514.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649-667.

Kang, S. H. K. (2010). Enhancing visuospatial learning: the benefit of retrieval practice. *Memory and Cognition 38*(8), 1009-1017.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review, 21*(6), 1544-1550.

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (J. H. Byrne, Series Ed.) (pp. 487-514). Oxford, United Kingdom: Academic Press.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*, 317-326.

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

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation, Volume 61* (pp. 237-284). San Diego, CA: Elsevier Academic Press.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250-1257.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966-968.

Kelly, J. W., Carpenter, S. K., & Sjolund, L. A. (2015). Retrieval enhances route knowledge acquisition, but only when movement errors are prevented. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1540-1547

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283-294.

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*(2), 291-304.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442.

Little, D. R., Lewandowsky, S., & Heit, E. (2006). Ad hoc category restructuring. *Memory & Cognition*, *34*(7), 1398-1413.

Logan, J. M., & Balota, D. A., (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging Neuropsychology, and Cognition, 15*(3), 257-280.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psycholo*gy, *38*(2), 94-97.

Mannes, S. M., & Kintsch W. W. (1987). Knowledge organization and text organization. *Cognition & Instruction*, *4*, 91-115.

McDaniel, M. A., & Butler, A. C. (2010). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175-198). New York, NY: Psychology Press.

McDaniel, M. A., Moore, B. A., & Whiteman, H. L. (1998). Dynamic changes in hypermnesia across early and late tests: A relational/item-specific account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(1), 173-185.

McDaniel, M. A., & Einstein, G. O., (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*(2), 113-145.

McDaniel, M. A., & Einstein, G. O., (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed), *Experimental cognitive psychology and its applications* (pp. 73-85). Washington, DC: American Psychological Association.

McDaniel, M. A., Howard, D. C., & Einstein, G. O., (2009). The read-recite-review study strategy: effective and portable. *Psychological Science, 20*(4), 516-522.

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review 27*(2), 327-331.

Rawson, K. A., & Kintsch, W. (2002). How does background information improve memory for text content? *Memory & Cognition, 30,* 768-778.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*(1), 45-48.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69*(4), 344-354.

Tulving, E. (1966). Subjective organization and effects of repetition in multi-trial free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 5*, 193-197.

van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1532-1541.

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247-264.

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review*, *27*(2), 265-289.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004) Category norms: An updated an expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289-335.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1036-1046.

Wissman, K. T., Rawson, K. A., & Pyc, M. A., (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140-1147.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995-1008.

APPENDIX A: TABLES

Table 1

Proportion of Idea Units Recalled in the Learning Phase of Experiment 1

Text Structure	Period 1	Period 2
Intact	.36 (.03)	.54 (.04)
Scrambled	.24 (.03)	.39 (.04)

Note. Standard errors of the mean are shown in parentheses.

Table 2

Proportion of Idea Units Recalled During Session 2 as a Function of Text Structure and Initial Learning Activity

Text Structure	Period 1	Period 2
Intact	.30 (.04)	.48 (.05)
Scrambled	.21 (.04)	.33 (.04)

Note. Standard errors of the mean are shown in parentheses.

Table 3

Proportion of Idea Units Recalled in the Learning Phase of Experiment 2

Topic Word	Period 1	Period 2
Present	.51 (.03)	.66 (.04)
Absent	.44 (.03)	.60 (.04)

Note. Standard errors of the mean are shown in parentheses.

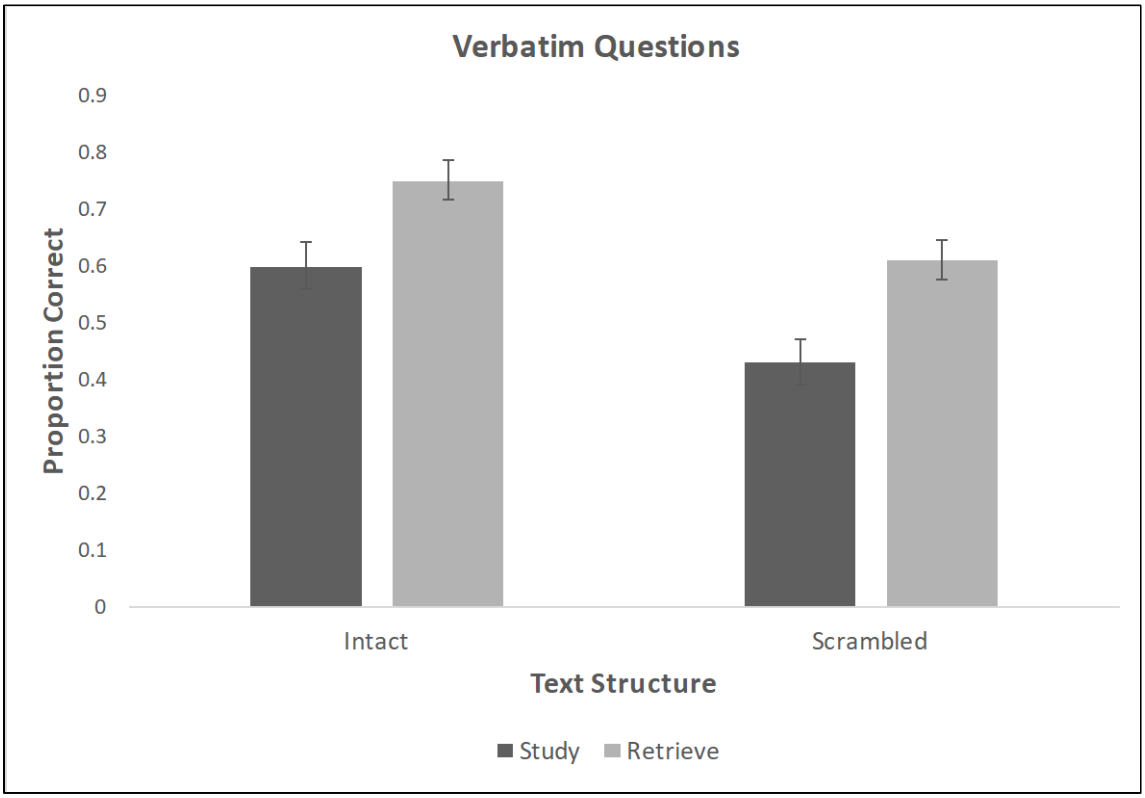
Table 4

Proportion of Words Recalled in the Learning Phase of Experiment 3

Category Names	Period 1	Period 2
Present	.53 (.03)	.71 (.03)
Absent	.51 (.04)	.72 (.04)

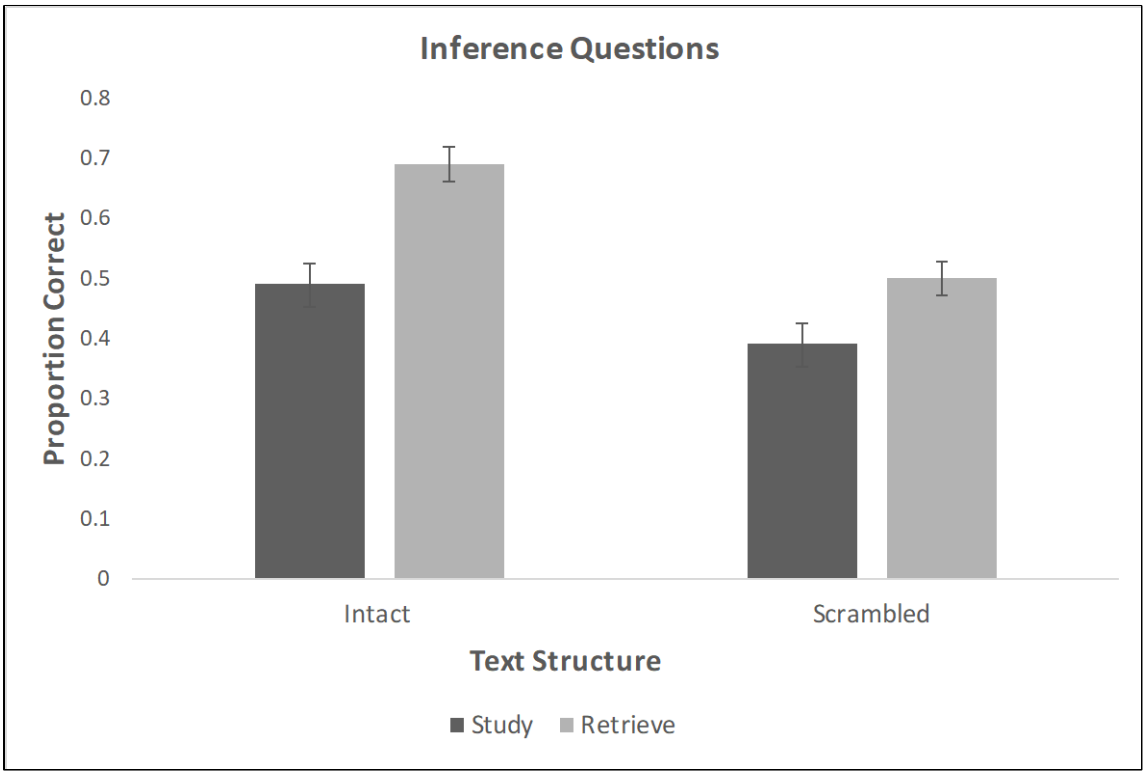
Note. Standard errors of the mean are shown in parentheses.

APPENDIX B: FIGURES



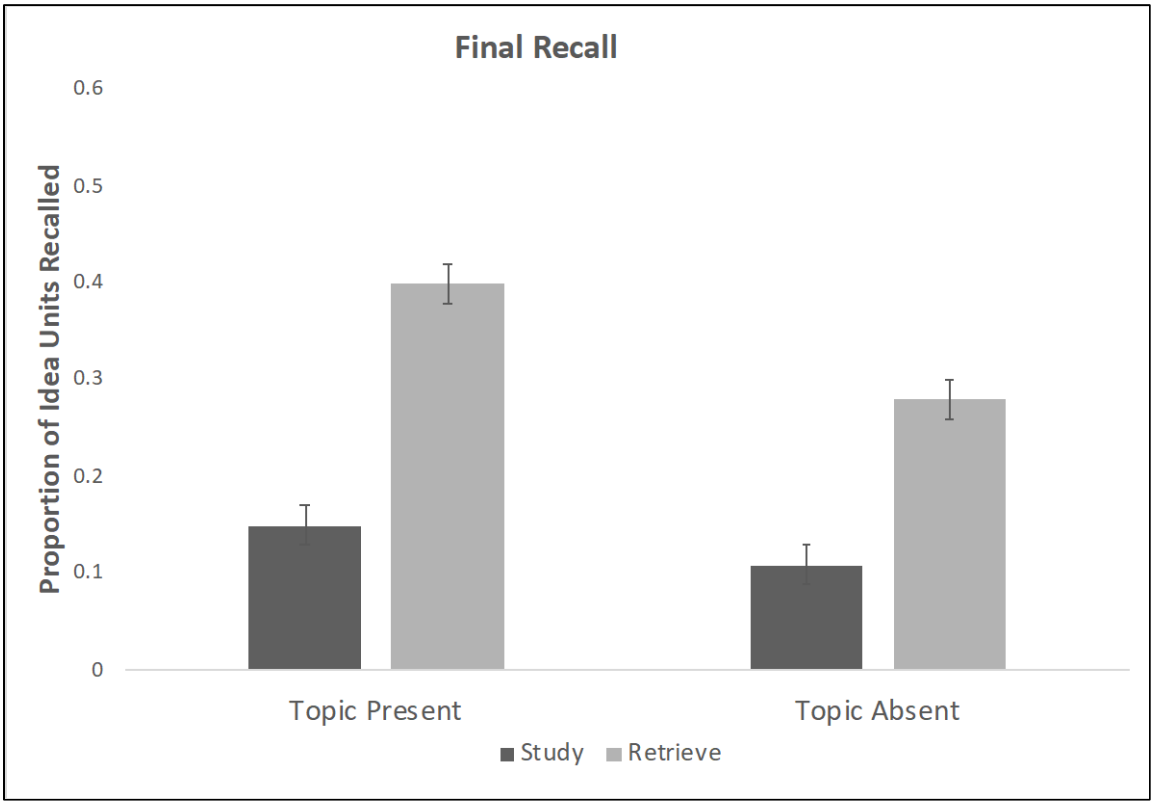
Note. Error bars represent standard errors of the mean.

Figure 1. Final short answer performance for verbatim questions.



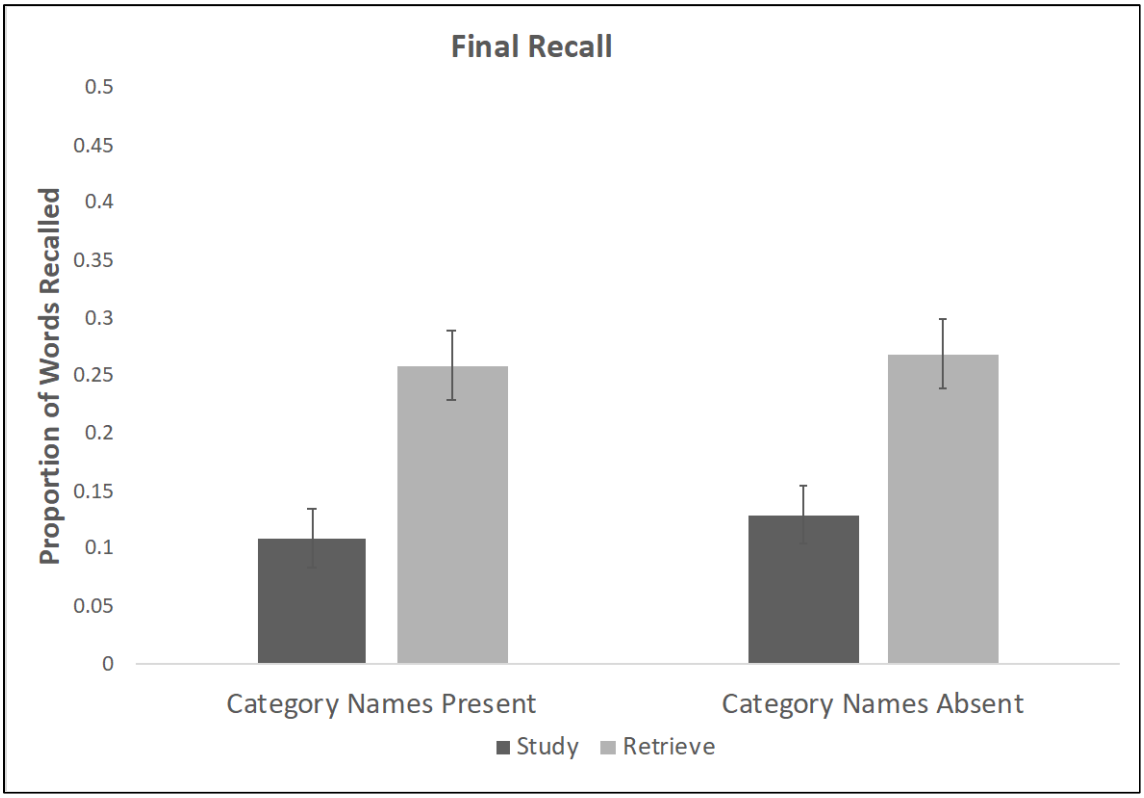
Note. Error bars represent standard errors of the mean.

Figure 2. Final short answer performance for inference questions. Error bars represent standard errors of the mean.



Note. Error bars represent standard errors of the mean.

Figure 3. Final recall performance in Experiment 2. Error bars represent standard errors of the mean.



Note. Error bars represent standard errors of the mean.

Figure 4. Final recall performance in Experiment 3. Error bars represent standard errors of the mean.

APPENDIX C: INTACT AND SCRAMBLED VERSIONS OF THE "TROPISMS" PASSAGE IN EXPERIMENT 1

Intact

Growing plants can respond to a stimulus coming from a given direction by growing more rapidly on one side than the other and hence bending toward or away from the stimulus. This growth response in plants is defined as tropism. Tropisms can occur only in those parts of the plant that are growing and elongating, such as the plant stem or root. For example, a plant leaf on the windowsill will gradually grow so that the stems bend toward the light source. The bending of the stems occurs because the cells on the nonlighted side grow more rapidly than those facing the light. The particular chemical responsible for this growth is called an auxin.

Tropisms are named for the kind of stimuli eliciting them. A phototropism is a growth response to light. The plant on the windowsill described above is a good example of a phototropic response. Geotropism is a growth response to gravity. The root of the plant is geotropic because it grows toward the force of gravity. Two other forms of tropism are chemotropism (a growth response to some chemical) and thigmotropism (a growth response to contact). Bean plants are famous for their thigmotropism. Once contact is made with the top of a bean stem, it curls, producing the clinging response typically found in these plants.

A tropic growth may be either positive (toward the stimulus) or negative (away from the stimulus). For example, a seed always grows with the root downward and the stem upward. Thus, the root is positively geotropic and the stem is negatively geotropic.

Scrambled

The bending of the stems occurs because the cells on the nonlighted side grow more rapidly than those facing the light. The root of the plant is geotropic because it grows toward the force of gravity. A plant on the windowsill is a good example of a phototropic response. The root is positively geotropic and the stem is negatively geotropic. Two forms of tropism are chemotropism (a growth response to some chemical) and thigmotropism (a growth response to contact). A seed always grows with the root downward and the stem upward.

The particular chemical responsible for phototropic growth is called an auxin. Once contact is made with the top of a bean stem, it curls, producing the clinging response typically found in bean plants. A tropic growth may be either positive (toward the stimulus) or negative (away from the stimulus). Geotropism is a growth response to gravity. A plant leaf on the windowsill will gradually grow so that the stems bend toward the light source. Tropisms can occur only in those parts of the plant that are growing and elongating, such as the plant stem or root. Bean plants are famous for their thigmotropism. A phototropism is a growth response to light.

Growing plants can respond to a stimulus coming from a given direction by growing more rapidly on one side than the other and hence bending toward or away from the stimulus. Tropisms are named for the kind of stimuli eliciting them. A growth response in which plants bend toward or away from a particular stimulus is defined as tropism.

APPENDIX D: INTACT AND SCRAMBLED VERSIONS OF THE "HOMEOSTASIS" PASSAGE IN EXPERIMENT 1

Intact

The human body has an amazing capacity to speed up or slow down physiological processes when changes occur in internal states. This ability is defined as homeostasis. The most sophisticated system in our body which carries out homeostasis is the endocrine system. This is a series of glands in our body that produce hormones. The endocrine system operates on a principle similar to a thermostat. A thermostat detects the need for heat, turns on the furnace when the temperature is too low, and then turns off the furnace when the temperature is again normal.

One example of homeostasis in action involves the hormone vasopressin. Vasopressin causes the capillaries to constrict, and when the body suffers severe bleeding due to an injury, the amount of this hormone is drastically increased. This helps to slow down blood flow by closing off small blood vessels. Thus, blood flow to the injured area is reduced. The antidiuretic hormone, ADH, helps the body conserve water by directing the kidneys to reabsorb water. A normal amount of ADH tells the kidneys to reabsorb all but one liter of water daily. However, when the body becomes dehydrated from water loss due to perspiration during hot weather, more ADH is released telling the kidneys to reabsorb more water than usual to make up for that loss.

Sometimes the production of a hormone in the body may be either overactive or underactive, regardless of internal needs. If it is overactive, it is called "hyper-" and if it is underactive, "hypo-". For example, hyperthyroid conditions produce too much growth while hypothyroid conditions produce stunted growth.

Scrambled

Blood flow to the injured area is reduced. The endocrine system is a series of glands in our body that produce hormones. When the body becomes dehydrated from water loss due to perspiration during hot weather, more ADH is released telling the kidneys to reabsorb more water than usual to make up for that loss. A thermostat detects the need for heat, turns on the furnace when the temperature is too low, and then turns off the furnace when the temperature is again normal. One example of homeostasis in action involves the hormone vasopressin. The most sophisticated system in our body which carries out homeostasis is the endocrine system.

The increase in vasopressin helps to slow down blood flow by closing off small blood vessels. Hyperthyroid conditions produce too much growth while hypothyroid conditions produce stunted growth. A normal amount of ADH tells the kidneys to reabsorb all but one liter of water daily. The human body has an amazing capacity to speed up or slow down physiological processes when changes occur in internal states. If the production of a hormone is overactive, it is called "hyper-" and if it is underactive, "hypo-". The endocrine system

operates on a principle similar to a thermostat. The ability to speed up or slow down physiological processes when changes occur in internal states is defined as homeostasis.

Vasopressin causes the capillaries to constrict, and when the body suffers severe bleeding due to an injury, the amount of this hormone is drastically increased. The antidiuretic hormone, ADH, helps the body conserve water by directing the kidneys to reabsorb water. Sometimes the production of a hormone in the body may be either overactive or underactive, regardless of internal needs.

APPENDIX E: "LAUNDRY" PASSAGE USED IN EXPERIMENT 2

The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavor. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many can easily arise. A mistake can be expensive as well. The manipulation of the appropriate mechanisms should be self-explanatory, and we need not dwell on it here. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell.

APPENDIX F: WORD LISTS USED IN EXPERIMENT 3

A thing that makes noise	A thing that is green	A thing made of wood
Airplane	Clothes	Tree
Drum	Frog	Floor
Gun	Lettuce	Paper
Baby	Plant	Door
Train	Pea	Pencil
Radio	Eyes	Dresser