The Effect of JOLs and Free Recall on Reading Comprehension and Study Choices

Tara L R Beziat[1] & Christopher A Was[2]

[1] Auburn University at Montgomery, USA

[2] Kent State University, USA

Correspondence: Tara L Romes Beziat, Assistant Professor of Educational Psychology, Auburn University at Montgomery, USA.

Received: November 15, 2015 Accepted: December 8, 2015 Online Published: December 15, 2015

doi:10.5430/irhe.v1n1p60 URL: http://dx.doi.org/10.5430/irhe.v1n1p60

Abstract

One way to improve students' academic performance is to improve their reading comprehension. Previous investigations demonstrated that testing students on the material as well as having them use metacognitive strategies have independently improved reading comprehension. The tests used in the learning phase in previous investigations have typically been experimenter created. In the current study, free recall of recently read text was used as the test in the learning phase. A second important aspect of the current investigation is the inaccuracy of students' meta-comprehension judgments. Although use of metacognitive strategies does improve academic performance, students often make inaccurate judgments about what they know and are particularly inaccurate in their assessment of text comprehension. The aim of this study was to determine if free recall was an effective testing strategy for reading comprehension and long- term retention. Finally, this study explored the relationship between judgments of learning and re-study choices. Retrieval practice, more precisely free recall, did improve the accuracy of judgments of learning in comparison to rereading the material. However, free recall did not improve participants' academic performance or long-term retention of material.

Keywords: metacognition, retrieval practice, study choices, expository text comprehension

1. Introduction

In many classrooms, reading is an integral part of learning and therefore educators have continued to explore ways to help children comprehend expository text. Different types of reading and note-taking strategies for the content areas have been developed to improve reading comprehension. Along with trying to improve reading comprehension during the initial learning phase, educators have also tried to improve the study habits of their students. Often students complain they do not know how to study and perform poorly on a test despite studying for hours. Changing students' perceptions of studying and helping them improve their study choices are key components to improving their learning.

This study examined the effects of testing, or more precisely, retrieval practice on academic preparation and performance. One goal of this research was to understand the effects of retrieval practice, in the form of free recall, on reading comprehension and long-term retention. To accomplish this goal, we examined whether retrieval practice was an effective learning method for use with an extensive complex reading selection. A second goal of the proposed research was to examine the relationship between retrieval practice and judgments of learning (JOLs). A key part of the learning process is accurately assessing the information that needs to be reviewed for greater understanding. One possible way to improve the accuracy of JOLs and the studying process would be to incorporate retrieval practice.

1.1 Testing Effect and Retrieval Practice

Research based on the work of Tulving (1967) and Roediger and Karpicke (2006) has continuously shown that testing is an effective learning tool. Indeed, it has been frequently demonstrated that retrieval practice is more effective than increased study time for long-term retention (see Roediger & Butler, 2011 for a review). Though tests, or acts of retrieving information, are proven learning tools students often rely on re-reading material and highlighting or underlining key words to study when presented with text. These methods for studying are less effective than repeated retrievals for future test performance (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013). Testing in

the form of retrieving information from long-term memory increases retention, enhances transfer of learning, (Roediger & Butler, 2011), facilitates the learning of functions (Kang, McDaniel, & Pashler, 2011), and facilitates the learning of new material (Wissman, Rawson, & Pyc, 2011). These previous studies demonstrate the efficacy of recall or retrieval practice as a study strategy.

1.1.1 Retrieval Practice with Word-Pairs

Much of the early work in testing effect research used the common methodology of presenting word-pairs, or word-lists, to the subjects for learning followed by free recall of as many words or word-pairs as possible (Kuo & Hershman, 1996; Karpicke &Roediger, 2007a; Hays, Kornell, & Bjork, 2010). Free recall has proven to be an effective way to learn simple materials such as word pairs. Kuo and Hershman (1996) found subjects were better able to recall words on the final test when they were tested during learning, compared to when they were simply asked to read words aloud. Karpicke and Roediger (2007a) examined whether repeated retrieval practice increased long-term retention of material. Participants in the retrieval practice, in the form of free recall, condition were able to recall more words after one week, than those in the re-reading condition. Their results indicated that free recall of words in the learning phase could improve long-term retention of the material.

Researchers have also used word-pairs as the target material to examine the effects of repeated retrieval. Word-pair learning mimics many classroom-learning situations. For example, students are often asked to study word pairs in their foreign language classes. In a study conducted by Hays, Kornell, and Bjork (2010) subjects were presented word-pairs (e.g. house-la maison). In the study phase, the subjects were presented the word-pair and instructed to learn it for a future test. In the testing phase, the subjects were presented one of the words (e.g. house-) and instructed to recall the other word. On the final test, subjects were again presented the retrieval prompt, house-, and instructed to recall the other word. Like the studies using lists of words, participants who were provided retrieval practice outperformed the subjects who repeatedly studied the material.

In each of these studies a common methodology was used, participants were presented words to learn. The use of free recall as retrieval practice proved to be an effective learning tool. Participants who freely recalled the words in the learning phase outperformed those who restudied the material on subsequent tests. The present investigation addressed the question of whether free recall is an effective testing method during study periods for long sections of text, not just words or word-pairs. Though learning words and word-pairs enables researchers to examine the specific effects of testing on learning, students are often expected to read text to learn the material for class. Despite the numerous studies that have examined the testing effect, one limitation in the extant literature is apparent: the paucity of studies conducted examining whether retrieval practice is an effective learning strategy for learning from substantial amounts of text.

1.1.2 Retrieval Practice with Text Passages

Investigations conducted by Roediger and Karpicke (2006) and Rawson and Dunlosky (2011) provide evidence that the testing effect can be used with passages of text. However, these studies have specific limitations. First, the length of text used in these studies is short (260 words and approximately 420 words in length) in comparison to the average length of textbook chapters for high school and college students. Second, the experimenters provided guided prompts for the recall sessions. Therefore, these investigations were unable to provide evidence that free recall during the learning phase is an effective learning strategy for longer passages of text.

A small number of studies did use extensive text to examine the testing effect. An investigation conducted by Butler (2010) increased the reading expectations of the participants by having them read six passages (approximately 1000 words in length) that contained difficult concepts. Despite an increase in the length of reading material, this study found repeated retrievals of the material were more effective than increased study time. However, the tests used during the learning phase for this study were created by the experimenters.

In order to mimic real world classroom experiences, some researchers have used textbook chapters as the reading material (Karpicke & Blunt, 2011; Roediger, Agarwal, McDaniel & McDermott, 2010; McDaniel, Argarwal, Huelser, McDermott, Roediger, 2011). McDaniel et. al. (2011) asked participants to read a chapter, and then take a pre-lesson quiz. After the pre-lesson quiz, the teacher presented the material from the chapter and gave the participants a post-lesson quiz. Finally, after a 24-hour delay participants were given a review quiz. A summative assessment was given after the unit was completed which contained quizzed and non-quizzed items. Results showed a clear effect of testing on retention of chapter concepts. Put differently, participants performed better on the quizzed items than the non-quiz items. Similar results were found in a study conducted in a social studies class as well (Roediger, Agarwal, McDaniel & McDermott, 2011).

The results of these investigations provided evidence that the testing effect transfers to longer passages of text like a textbook chapter. However, in each of these experiments the experimenters created the testing material. The "testing effect" was created by administering an experimenter created exam of the to be learned material, not by requiring participants to freely recall the content. This leaves the unanswered question of whether the testing effects will be present when the participants are expected to test themselves by freely recalling the information they read. One aim of the present study was to examine if free recall is an effective testing strategy for extensive reading passages.

1.2 Metacognition and Judgments of Learning

Metacognition is described as thoughts about one's knowledge and control over one's cognitive processes (Flavell, 1979). Research has shown a positive relationship between metacognitive practices and learning (see Finley, Tullis, & Benjamin, 2010, for a review). Schmitt and Newby (1986) recognized that metacognitive strategies needed to be incorporated into instruction. Building these components into instruction helps the learner regulate the processing of information but also ensures they have a better understanding of the information. A key part of regulating learning is monitoring the inflow of information and taking actions when the information does not match our previous understandings (Nelson & Narens, 1990).

When JOLs are used in experimental tasks, participants are asked to examine how well they have learned recent material and make a judgment on their ability to recall the information at a later time. For example, participants are shown an English-French word pair (house -- la maison). After studying the word pair, they are asked, how likely they are to remember the English translation, when they are given the French word. They make a judgment, of 0% to 100% chance of remembering. Initial research showed individuals often are inaccurate in their assessment of what they know and do not know (Maki & Berry, 1984). For example, participants often provide high estimates of the chance they will remember the word or to be learned information, but then on the test they do not recall the necessary information. Said differently, they are not able to predict with accuracy what they will and will not remember on future academic performances. Maki and colleagues consistently found low correlations between prediction judgments and performance, specifically on measure with between reading comprehension and metacomprehension (Maki & Berry, 1984; Maki, Foley, Kajer, Thompson & Willert, 1990; Maki, Jonas & Kallod, 1994). Said differently, participants do not accurately predict how they will do on a future performance.

The poor calibration between prediction judgments and performance has been frequently demonstrated. However, Nelson and Dunlosky (1991) and Maki (1998) were able to improve the accuracy of judgments of learning by having students make delayed judgments of learning versus immediate. If one has to attempt to recall the information from long-term memory to make a judgment of learning, their accuracy increases for judging future performance because they have attempted to recall the information, rather than simply access information in working memory.

Having participants reread the text or summarize the text can also create delays. Rawson, Dunlosky and Theide (2000) investigated whether rereading text improved participant's metacomprehension. In both experiments, those who reread the text outperformed the control group on a subsequent multiple choice test and were better able to predict how well they would do on the multiple choice test. Thiede and Anderson (2003) found the accuracy of judgments can be improved through delayed summarizing of the text and generating key words. The delay between reading and summarizing ensured that individuals were accessing information from their long-term memory. When individuals assessed future performance based on their ability to retrieve information from long-term memory the accuracy of their judgments improved.

1.3 Study Choices

A third aim of this investigation was to examine the relationship between judgments of learning (JOLs) and re-study choices. An integral part of the learning process is knowing what you know and using this information accordingly. However, people often inaccurately judge what they do and do not know and make inaccurate judgments about future performances (Kruger & Dunning, 1999). These inaccurate judgments affect how individuals regulate their learning and self-regulation is a key component in learning (Whitebread & Pino-Pasternak, 2010). In order to allocate adequate time to study, one needs to regulate their learning.

Two hypotheses on the allocation of study time and academic performance are the Discrepancy Reduction Model (Thiede &Dunlosky, 1999) and Region of Proximal Learning Hypothesis (Metcalfe, 2002). Thiede and Dunlosky (1999) stated students allocated study time based on the perceived difference between what they know and what they have to know and attempt to reduce this difference. Yet, there is some evidence to show that if this discrepancy is too large and there is a time constraint, students focus on the easier items (Son & Metcalfe, 2000). The Region of Proximal Learning hypotheses (Metcalfe, 2002) proposes that students prefer items of easy and medium difficulty, because they

provide access to the most efficient gains. Only as the items of easy and medium difficulty are learned, do they turn to the truly difficult items.

The Discrepancy Reduction Model and the Region of Proximal Learning hypotheses leave gaps in the understanding of how students allocate study time. At times students may study the difficult items first, at other times they may study the easy items first. Yet, the premise of both of these hypotheses is that learners create an agenda for studying based on the difficulty of the items. However, recent evidence suggests that there are other reasons participants in experimental settings may choose the items to study that they do. For example, Ariel, Al-harthy, Was, and Dunlosky (2011) postulated that it is our reading habits that influence study choices not item difficulty. Their reading habit bias, or the direction in which they read influenced the learners' decisions for study choices. Native English readers chose to study the word-pair on the left and native Arabic readers chose to study the word-pair on the right, despite the level of difficulty of the item.

Students in the current study made judgments of learning and were able to select restudy choices while viewing their judgments of learning. If individuals select to restudy segments to which they gave higher judgments of learning (meaning they understand the material), this would provide evidence for Metcalfe's model. However, if they select those segments they gave low judgments of learning to (meaning they did not understand the material), this would provide evidence for Thiede and Dunlosky's model. If participants select the first items and there is no relationship between the JOLs and the study choices, this would provide evidence for influence of reading habit bias.

Reading comprehension and long-term retention are keys to academic success. Research overwhelmingly shows retrieval practice is more effective than restudying for learning. Although creating self-made tests may prove difficult or time consuming for the average student, incorporating simple retrieval of information from memory without specific questions or prompts would be less labor intensive for students. Therefore retrieval practice may be an effective study method and one that students can easily implement. There is a large body of research, which shows a positive relationship between metacognitive strategies and learning. Yet, individuals' predictions about future performance are often inaccurate. The current study attempted to answer the following questions: 1) Does free recall after reading improve reading comprehension and long-term retention of material? 2) Does free recall after reading improve the accuracy of JOLs? 3) Is there a relationship between JOLs and study choices?

2. Methods

2.1 Participants

One hundred and five undergraduates at a large mid-western university who were enrolled in educational psychology classes participated in the current study. Participants received course credit for their participation. The mean age was 20.5 years and 81 percent of those who took part in the study were female.

2.2 Materials

Texts: Two passages of texts were used in the study. The reading passages for this experiment were adapted from textbooks used in introductory college courses at another university representing the domains of history and economics. The passage, "The Economics of Wealth" had a Flesh-Kincaid readability score of 11.3 and the passage "Catherine the Great" had a Flesh-Kincaid readability score of 11. Each passage consisted of 6 segments and each segment containing 400-600 words, with the total passage being approximately 3000 words.

2.3 Procedures

The experimental procedures were conducted in a well-lit room with four sound dampening computer carrels each with a personal computer with a 19" VGA monitor.

All participants were instructed that they were to read the presented texts with the intent to comprehend the material. The instructions also explained that the passages were divided into multiple segments. Each segment was numbered because they would be given an opportunity at the end to select half of the segments to reread and review before they answered questions based on the text. The instructions also included that the participants may be asked to make judgments about their learning and may be asked to recall information after they read a segment of the text.

In all conditions participants performed a reading comprehension test presented via computer. After reading the instructions, all participants were presented with the title of the passage and the first segment of the passage. Participants had unlimited time to read each segment. When ready to move on, participants pressed the spacebar. In the recall condition, participants were then asked to recall as much information as possible about the previous segment. After pressing the spacebar, a free recall screen appeared and participants were instructed to type as much as they could remember from the previous segment in a space provided on the computer screen. The participants

were instructed to press '9' when they were finished with the free recall screen.

Participants in the study condition were also given unlimited time to read each segment of the passage. However, participants in the study condition were not asked to freely recall as much as they could after completing their first reading of a segment, but instead were presented with same segment and instructed to study the passage as long as necessary until they felt that completely understood the segment. When ready to move on they were instructed to push the spacebar.

In both conditions, the next step was to make a segment JOL. After the free recall screen/restudy screen, participants were asked to make a JOL. Participants were presented with the question, "How well do you understand the information you have just read?" Participants rated their understanding on a scale from 0, "I don't understand any of the information I just read" to 10 "I understand all of the information I just read." After participants made a JOL they moved on to the next segment.

After the entire passage was read, participants in both the study and recall conditions were asked to select three of the six segments they would like to review. Their JOLs for each segment were presented on the computer for them to reference. Participants used the keyboard to indicate the three segments they would like to restudy. In the honor condition, the computer presented the three segments they chose, one at a time for self-paced study. However, in the dishonor condition participants were presented with the segments they did not choose, one at a time for self-paced study. For example, if a participant selected to restudy segments one, two, and three, they were presented with segments four, five and six.

After the restudy phase all participants completed the immediate test. The immediate test was comprised of 6 open-ended questions (one per passage segment) and 12 multiple-choice questions (two questions related to each of the six passage segments) designed to evaluate their understanding of the passage. Per segment, one multiple-choice question was about a specific detail in the passage and one multiple-choice question was an inferential question. After answering all questions, students saw a screen to tell the researcher they had completed the study. Students were asked to return one week later to complete the same test of 6 open-ended questions and 12 multiple-choice questions.

3. Results

A total of 105 subjects took part in the current study. However, 25 were excluded in the analysis (unless otherwise specified) because they did not complete all parts of the immediate or delayed test. Table 1 presents the means and standard deviations for immediate and delayed test scores by condition. The two tests used in the current study, the immediate and delayed test, were identical. Each contained the same 12 multiple-choice questions (Catherine Test Cronbach *alpha*=.636; Economics Test Chronbach *alpha*=.510) and 6 open-ended questions.

Table 1. Means and standard deviations for immediate test scores and delayed test scores

Source	*M*	*SD*
Immediate Test		
Test Condition	1.023	.367
Study Condition	1.014	.370
Honor Condition	1.049	.364
Dishonor Condition	.998	.370
Economy Text	.912	.340
Catherine the Great Text	1.114	.366
Delayed Test		
Test Condition	.966	.375
Study Condition	.934	.345
Honor Condition	.946	.366
Dishonor Condition	.953	.355
Economy Text	.887	.358
Catherine the Great Text	1.007	.353

Three raters, who are current or previous social studies teachers, graded the open-ended questions using a rubric. Raters graded each response using a 0-2 scale. Correct answers were assigned a value of 2. If the answer given was

partially correct a score of 1 was given. If no answer or a completely inaccurate answer was given a score of 0 was specified. A total of 537 responses for the Catherine the Great reading were analyzed. The intraclass correlation coefficient (ICC) was .75, 95% CIs [.69, .79], α=.91, for the three raters, indicating agreement. For the Economy text a total of 500 responses were analyzed. The ICC was.75, 95% CIs[.71, .78], α =.90, again indicating agreement. For each test, analyses were conducted on the multiple-choice questions, open-ended responses and a combined accuracy score composed of the scores on open-ended and multiple-choice questions.

3.1 Immediate Test Results

Data were screened to ensure that the assumptions of a factorial ANOVA were fulfilled. Participants with missing test scores were eliminated. An examination of outliers revealed three outliers in the immediate test multiple choice scores for the economy test with scores of 1.00 (subjects 64, 42, and 3). These outliers were not removed from analysis because these participants represented participants who successfully completed the immediate multiple-choice questions. A total of 80 cases were used in future analyses. Next, an ANOVA was conducted to determine if there was an effect of text on total scores for the immediate and delayed tests. Results indicated a statistically significant difference for the total scores on the immediate test between the two texts, $F(1,79)=6.48$, $p = .013$ (See Table 2 for a summary of results).

Table 2. One-way ANOVAs comparing texts

Source	*SS*	*df*	*MS*	*F*	*p*
Immediate Test Scores	.806	1	.806	6.370	.014
Delayed Test Scores	.280	1	.280	2.201	.142

Based on these results, text was included in the factorial ANOVAs for the immediate tests to determine if there were any interactions.

A 2(reading condition: reread vs. recall) x 2(JOL condition: honor vs. dishonor) x 2(text; economy or Catherine the Great) factorial analysis of variance was conducted. A summary of the results are presented in Table 3.

Table 3. A factorial ANOVA for immediate total scores

Source	*SS*	*df*	*MS*	*F*	*p*	*partial η²*
ReadingCondition	.003	1	.003	.019	.890	.000
JOLCondition	.069	1	.069	.526	.471	.007
Text	.798	1	.798	6.113	.016	.078
ReadingCondition * JOLCondition	.028	1	.028	.216	.644	.003
ReadingCondition * Text	.013	1	.013	.102	.750	.001
JOLCondition * Text	.011	1	.011	.087	.769	.001
ReadingCondition * JOLCondition * Text	.251	1	.251	1.921	.170	.026

Main effect results revealed that immediate test scores were not statistically different between participants in the different reading conditions (recall vs. reread). Also, immediate test scores were not statistically significant between participants in the JOL conditions (honor vs. dishonor). The partial eta squared for the three conditions is minute revealing that each condition had little effect on the immediate test scores. Results indicate no significant interactions between the three independent variables.

3.2 Delayed Test Results

Analyses showed no statistically significant differences between those who read the economy passage and those who read the Catherine passage on delayed test scores. Therefore, text was not included as a factor in the analysis. A 2(reading condition: reread vs. recall) x 2(JOL condition: honor vs. dishonor) analysis of variance was conducted for the delayed test scores; a summary of the results are presented in Table 4.

Table 4. A factorial ANOVA for delayed test score

Source	*SS*	*df*	*MS*	*F*	*p*	*Partial η²*
Reading Condition	.020	1	.020	.154	.696	.002
JOL Condition	.001	1	.001	.006	.937	.000
ReadingCondition * JOLCondition	.004	1	.004	.032	.859	.000

The interaction between reading condition and JOL condition was not significant. Also, neither main effect was found to be significant. As with the immediate test scores, recall did not prove to be more effective than restudying and participants whose restudy selections were honored did not outperform those whose restudy selections were dishonored.

3.3 Comparison of the Immediate and Delayed Test Scores

First, a paired-samples t-test was conducted to see if there was a difference between the immediate test scores and delayed test scores. The results revealed there was a statistically significant difference between the set of test scores, $t(79)= 3.34$, p=.001. The examination of test score means revealed individuals had higher test scores on the immediate test (M=1.014, SD=0.367) than on the delayed test (M=0.949, SD=0.359).

Next, a repeated measures ANOVA was conducted to see if these differences over time were a result of the reading condition or JOL condition. Results from the multivariate tests indicate a statistically significant effect for the within-subjects factor, test, Wilk's $\Lambda = .87$, $F(1,76)=11.36$, $p<.001$, $m\eta^2=.13$. Nonetheless, there were no statistically significant interactions between the test and the different conditions (see Table 5 for a summary of results).

Table 5. Repeated measures ANOVA for immediate and delayed tests

Source	df	F	p	Partial η²
Test	1	11.356	.001	.130
Test * Reading Condition	1	.180	.672	.002
Test * JOL Condition	1	3.242	.076	.041
Test * Reading Condition * JOL Condition	1	1.140	.289	.015

This indicates that the changes in test scores do not change across the different conditions. Put differently, the test scores did not change as a result of the reading condition or the JOL condition. Further analysis revealed performance decreased over time for all conditions.

3.4 Judgments of Learning

To determine if participants JOLs affected their selections for restudy, participants' JOL scores were correlated to their restudy selections. Participants made a total of six JOLs, with scores ranging between 0-10. They were able to select three segments to restudy. For analysis, the segments participants selected to restudy were coded "1" and their non-selections were coded "0." For this analysis all 105 participants were used. JOL scores were negatively correlated with restudy selections for three segments (See Table 6 for a summary of the correlations).

Table 6. Correlations between segment JOLs and restudy selection

Source	***r***	***p***
Segment 1	-.213*	.029
Segment 2	-.248*	.011
Segment 3	-.139	.157
Segment 4	.101	.305
Segment 5	.091	.357
Segment 6	-.241*	.013

Note: *indicates statistically significant at the .05 level.

When participants gave the segment a lower JOL they were more likely to select that segment to restudy. Participants chose to restudy material they felt they did not learn well.

To further analyze these results, participants were organized into two groups using a mean split based on participants' JOLs for each segment. A low JOL group and a high JOL group were created per segment. For each segment, except segment 4, participants who provided a lower JOL were more likely to select that segment for restudy than those who provided higher JOLs (see Table 7). When individuals gave a segment a lower JOL they were more likely to select it for restudy.

Table 7. Mean split of JOLs

Source	Low JOL Avg	Average Selection	High JOL Avg	Average Selection
Segment 1	3.905	0.604	7.154	0.404
Segment 2	4.057	0.566	7.423	0.346
Segment 3	4.000	0.547	7.558	0.442
Segment 4	4.038	0.509	7.077	0.538
Segment 5	3.830	0.585	7.154	0.442
Segment 6	4.057	0.566	7.019	0.442

Further analysis was conducted to examine the JOL scores by reading condition. Mean JOL scores were computed for each segment. Participants in the study condition consistently had higher JOLs per segment (See Figure 1).

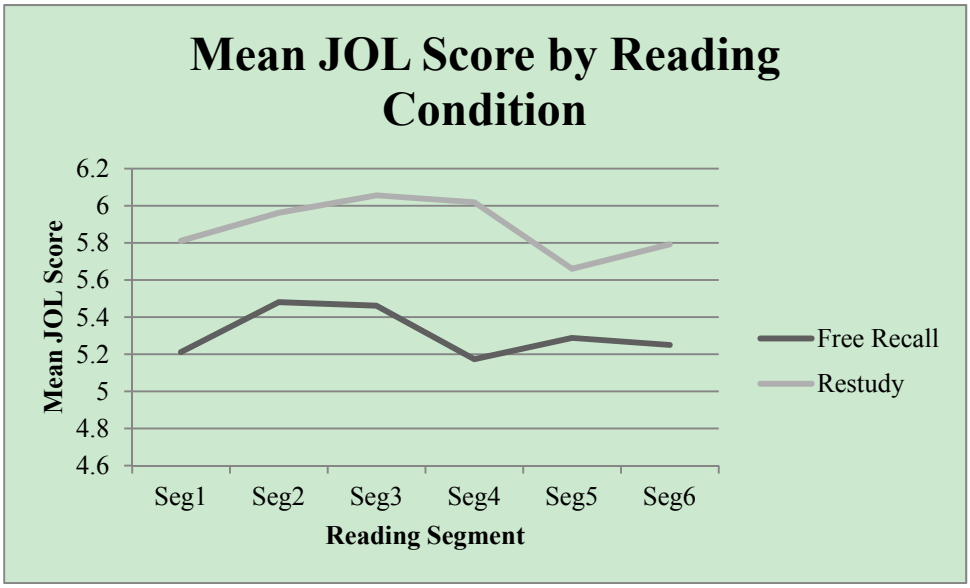

Figure 1

An independent samples t-test was conducted to see if there was a statistically significant difference between these means. Segment 4 had the only statistically significant difference; $t(103)= -2.41$, $p=.018$.

The final research question asked whether free recall after reading improved the accuracy of JOLs. Bias, a score that differentiates between over and under-confidence, was used to measure accuracy. Bias was calculated for the immediate and delayed tests using results from the multiple-choice and open-ended questions. Independent samples t-tests were conducted to see if there were differences between the test and study conditions (See Table 8).

Table 8. Bias differences between reading conditions

Source	*t*	*df*	*p*	95% CI
Immediate Test	-2.46	67	.016	[-2.17,-.25]
Delayed Test	-2.64	57	.011	[-2.33,-.32]

Statistically significant differences in bias were found between the free recall and study conditions for both the immediate, $t(67) = -2.46$, $p=.016$, and delayed tests, $t(57)= -2.64$, $p=.011$. Those in the study condition exhibited overconfidence in comparison to those in the test condition.

4. Discussion

A body of research has found that testing or more precisely repeated retrieval practice, improves future academic performance. However, in the current study testing did not prove to be more effective than restudying in regard to academic performance. Testing did not improve performance on the immediate or the delayed test. In previous studies, those in the testing condition performed similarly to those in the study conditions on immediate tests. Yet, on the delayed tests those in the testing condition outperformed those in the study condition. The current study did not replicate these findings.

One difference that distinguishes the current study from previous investigations was the length of the text. The current study used text that was close to 3000 words in length, whereas texts used in other studies ranged from 200 to 1000 words in length. Also, previous studies that used longer texts (e.g. a chapter or 1000 words) used an experimenter created test (e.g. multiple choice test) during the learning phase. In the current study, the method of testing during the learning phase was free recall. For each segment, which ranged in length of 400 to 600 words, participants were asked to type as much as they remembered from the segment. Considering the results of the current study it seems that when individuals are asked to read lengthier text in one sitting, similar to chapters in high school or college textbooks, free recall is not an effective testing method. Rather, individuals may need some type of structured quiz during the learning phase that provides immediate feedback about their learning (Pennebaker, Gosling & Ferrell, 2013). Pennebacker, Gosling and Ferrell (2013) found that when students received immediate feedback about their learning through weekly quizzes, this led to an improvement in their courses grades. It also lead to an improvement in their grades in other courses that semester and their GPA in the following term. Free recall may need to be paired with feedback to produce visible learning.

An additional part of the current study investigated if there was a relationship between JOLs and segments selected for restudy. The correlations revealed mixed results. Judgments of learning made for three of the segments, one, two and six, were negatively correlated with study choices. Put differently, when individuals gave segments lower JOLs, meaning they felt they did not understand the material, they in general opted to restudy that selection. When a low JOL group and a high JOL group were created per segment, those giving that segment a low JOL chose that segment for restudy more often than those who gave the segment a high JOLs. When individuals judged they did not understand a segment they generally chose to restudy that selection.

The results from this study provide some support for the discrepancy reduction model. Participants chose to restudy the lower rated segments. The discrepancy reduction model states that individuals try to reduce the larger discrepancies in their learning. They will choose to study the material they feel they do not understand. As previously stated, the correlations were not strong and only three were significant. However, there was a trend between lower JOLs and selection for restudy but more analysis is needed. Future research could focus on the relationship between JOLs and restudy selections using extensive text as in the current study. Of particular interest would be if the reading habit bias affects study choices when the target material is expository text.

A final goal of the current study was to examine whether or not free recall as a studying technique improved the accuracy of participants JOLs. Results of this study show that free recall reduces the confidence of individuals and therefore, improves the accuracy of their judgments. Participants who were able to re-read the material overall had higher JOLs, meaning they felt they knew the material. This could be a result of the fluency effect (Koriat & Ma'ayan, 2005). Koriat and Ma'ayan (2005) found when study time increased so did JOLs. Those who freely recalled the information were less confident in their knowledge of the material and this could have been a result of a lack of fluency in retrieval.

A key goal of metacognitive training has been to reduce overconfidence. Dunlosky and Rawson (2011) state, "The bottom line is that judgment accuracy matters a great deal for effective learning and durable retention" (pg.22) and

speculate that overconfidence leads individuals to prematurely stop studying and this may reduce retention of the material. In this particular study, reducing the participants' confidence did not lead to better academic performance. This is of interest because the participants could see the segments they felt ill-prepared for and chose to restudy those segment. The participants appeared to be making the appropriate choices but their actions did not lead to better test scores.

5. Limitations

The reading materials and tests used in the current study could be a possible limitation. The materials were screened ahead of time by outside reviewers for readability. In addition each text was analyzed using the Flesh-Kincaid scale. Also, outside reviewers compared the questions to the text to see if they matched. Originally, there were four sets of texts with corresponding tests and the two selected were the best of the four. Despite these precautions to make sure the materials were readable and the tests asked appropriate questions based on the text, more piloting of the data could have been done with the target population.

Another limitation and the most likely contributor to the null effects is a potential lack of motivation. The initial session in which participants were given the reading material and were administered the first test, often lasted close to an hour. Participants may have been more motivated to complete the task than to accurately answer the questions. Also, besides earning course credit and possibly personal satisfaction, students may have lacked motivation for completing the study. This lack of motivation could have led to sub-par performances. For example, when some individuals were given the opportunity to re-study the material they sometimes spent less than 30 seconds re-reading 500 words. Also, for the open-ended questions some of the missing data could be a result of individuals just skipping over a question. On the other hand, these behaviors mimic what students do in real learning situations when they reach a point of cognitive over load or sheer frustration.

6. Implications for Education and Future Research

Future research needs to continue to explore the possible limitations of the testing effect as demonstrated by the current study. The current study found that a combination of lengthy text and free recall did not lead to better performance. A future study should examine if the limitations on the testing effect are a result of this combination or one of the factors. For example, a study could have participants read a textbook chapter (at least 3000 words in length) that is divided into sections. After each section, one group could again freely recall the information. Another group could take a multiple-choice test, while a third group could answersopen- ended questions based on the text. This will allow researchers to better understand which testing methods work for materials that are used in everyday classrooms.

Researchers have begun examining the testing effect in classroom settings and have found positive results (Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Capenter, Pashler & Cepeda, 2009; McDaniel, Argarwal, Huelser, McDermott, & Roediger, 2011). In an attempt to determine if retrieval practice worked in a simulated classroom, Butler and Roediger (2007) conducted an investigation that used a simulated classroom experience and found retrieval practice to be effective. McDaniel et. al. (2011) and Roediger et al. (2010) found similar results in real world science and social studies classrooms. In each experiment, students were given multiple quizzes on the material from the textbook and teacher's lectures. Students performed better on the repeatedly quizzed material. Based on this line of research, if teachers want their students to know the material they have to quiz the students often on the material.

A key part of the learning process is monitoring one's learning and selecting what to restudy. Considering the current study, when individuals rate something with less confidence they are more likely to select that material to restudy. Yet, when these choices were honored this did not lead to better academic performance. There appears to be a flaw in the monitoring and control processes. It seems individuals should not be left on their own to make these restudy choices (Karpicke, 2009). Nonetheless, individuals cannot always access an algorithm that will analyze their studying and make the optimal choices for them. What can be done to help students make better study and restudy choices? Evidence has shown that individuals can be trained to better monitor their learning and reduce overconfidence. Was, Beziat, Isaacson (2013) found training in metacognitive activities over a semester, improved the knowledge monitoring of college students. Overall, a significant difference was seen between the students' pre-knowledge monitoring assessment scores and their post knowledge monitoring assessment scores. Students more accurately monitored their knowledge at the end of the semester after they had received training in metacognition.

When students can accurately judge what they do and do not know, they can better regulate their study time and choices. This being said, practice leads to better accuracy and less overconfidence. Therefore, teachers need to

provide opportunities for their students to practice monitoring their knowledge and seeing the consequences of this monitoring. Through this process, students can learn to gauge their learning and make better study choices for improved academic performance

References

Ariel, R., Al-Harthy, I. S., Was, C.A., & Dunlosky, J. (2011). Habitual reading biases in the allocation of study time. *Psychonomic Bulletin and Review, 18*, 1015-1021. http://dx.doi.org/10.1037/a0023064

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36,* 1118-1133. http://dx.doi.org/10.1037/a0019902

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527. http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760-771. http://dx.doi.org/10.1002/acp.1507

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2010). Metacognitive Control of Learning and Remembering. In M.S. Khine and I.M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp.109-131). New York, NY: Spring Science+Business Media.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin and Review, 17*, 797-801.

Kang, S. H. K, McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18,* 998-1005. http://dx.doi.org/10.3758/s13423-011-0113-x

Karpicke, J. D. (2009). Metacognitive Control and Strategy Selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469-486. http://dx.doi.org/10.1037/a0017341

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772-775. http://dx.doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L., III. (2007a). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162. http://dx.doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., III. (2007b). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 33,* 704–719. http://dx.doi.org/10.1037/0278-7393.33.4.704

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *The Journal of Memory of Language*, *52*, 478-492.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology, 109*, 451-464. Retrieved from http://www.jstor.org/stable/1423016

Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory and Cognition, 26*, 959-964. http://dx.doi.org/10.3758/BF03201176

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 663-679. http://dx.doi.org/10.1037/0278-7393.10.4.663

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609-616. http://dx.doi.org/10.1037/0278-7393.16.4.609

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1,* 126-129. http://dx.doi.org/10.3758/BF03200769

McDaniel, M. A., Agarwal, P. K., Huesler, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning

in middle school science classroom: The effects of quiz frequency and placement. Journal *of Educational Psychology, 103*, 399-414. http://dx.doi.org/10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494-513. http://dx.doi.org/10.1080/09541440701326154

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349-363. http://dx.doi.org/10.1037/0096-3445.131.3.349

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) Are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect". *Psychological Science, 2*, 267-270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp 125-173). New York: Academic Press.

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily Online Testing in large classes: Boosting college performance while reducing achievement gaps. *Plos ONE, 8*(11), e79774. http://dx.doi.org/10.1371/journal.pone.0079774

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?. *Journal of Experimental Psychology: General*, *140*(3), 283-302. http://dx.doi.org/10.1037/a0023956

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory and Cognition, 28*, 1004-1010. http://dx.doi.org/10.3758/BF03209348

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test- enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382-395. http://dx.doi.org/10.1037/a0026252

Roediger, H. L., III., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20-27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Sciences, 17,* 249-255. http://dx.doi.org/0.1111/j.1467-9280.2006.01693.x

Schmitt, M. C., & Newby, T. J. (1986). Metacognition: Relevance to instructional design. *Journal of Instructional Development,* 9, 29-33. http://dx.doi.org/10.1007/BF02908316

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129–160. http://dx.doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, Cognition, 25,* 1024-1037. http://dx.doi.org/10.1037/0278-7393.25.4.1024

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning Verbal Behavior, 6,* 175-184. http://dx.doi.org/10.1016/S0022-5371(67)80092-6

Was, C. A., Beziat, T. L. R., & Isaacson, R. M. (2013). Improving Metacognition in a College Classroom: Does Enough Practice Work?. *Journal of Research in Education, 23*(1), 77-93.

Whitebread, D., & Pino-Pasternak, D. (2010). Metacognition, Self-Regulation & Meta-Knowing. In Littleton, K., Wood, C. & Kleine Staarman, J. (Eds.), *International Handbook of Psychology in Education* (pp. 673–711). Bingley, UK: Emerald.

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,* 1140-1147. http://dx.doi.org/10.3758/s13423-011-0140-7