2013-08-27

Learnin' 'bout my generation: The effects of generation on encoding, recall, and metamemory

Burnett, Andrea

Burnett, A. (2013). Learnin' 'bout my generation: The effects of generation on encoding, recall, and metamemory (Unpublished doctoral thesis). University of Calgary, Calgary, AB. doi:10.11575/PRISM/26739
<http://hdl.handle.net/11023/887>

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Learnin' 'bout my generation:

The effects of generation on encoding, recall, and metamemory

by

Andrea Nicole Burnett

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

CALGARY, ALBERTA

AUGUST, 2013

© Andrea Nicole Burnett 2013

Abstract

My dissertation examined how encoding strategies, recall, and metamemory shift across two study-test experiences. Differential recall of generate targets and read targets on Test 1 led participants to develop an improved encoding strategy for their more poorly recalled target type, thus eliminating differential recall on Test 2 (Experiment 1-3). However, recall also improved across tests for groups that were not tested on both target types on Test 1 (Experiment 2), and for groups that studied and recalled only one target type (Experiment 1). Participants' reported strategies (Experiment 1) and metamemory judgments (Experiment 3) were used to elucidate how and when people modify their encoding strategies in an effort to improve future memory performance. Overall, the present study confirmed that people can learn about the effectiveness of a study strategy both during studying and on a test, and revealed that this learning is more ubiquitous and varied than previous research suggested.

Keywords: generation; encoding; recall; metamemory; strategies; study-test blocks

Acknowledgements

First and foremost, I would like to thank my adoring family for all of their love and support, through all of the sweat and tears that went into this project. Second, I would like to thank my supervisor, Dr. Glen Bodner, for his support, time, and patience. Third, I thank Sara Davis for her assistance with data collection. And finally, I thank my lab mates for providing me with distractions when I needed them.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Tables.....	vi
List of Figures.....	vii
List of Abbreviations and Nomenclature.....	ix
Introduction	1
Experiment 1	10
Method	12
Participants	12
Materials	12
Procedure	13
Results and Discussion	14
Recall in the within and between groups	15
Recall in the within group subgroups	18
Self-reported study strategies.....	23
Experiment 2A	29
Method	29
Participants	29
Materials and Procedure	30
Results and Discussion	30
Recall within the within group	30
Recall in the within group subgroups	31
Test 2 recall as a function of the Test 1 experience	33
Experiment 2B	35
Method	35
Participants.....	35
Materials and Procedure	35
Results and Discussion	35

Experiment 3	37
Method	39
Participants	39
Materials and Procedure.....	39
Results and Discussion	39
Recall within the within group.....	40
Recall in the within group subgroups	41
Combined Subgroup Analysis	43
Metamemory Measures	45
Metamemory in Block 1	45
Metamemorial changes across blocks	50
General Discussion	51
Accounting for the Elimination of the Generation Effect.....	52
Use of Study Strategies	57
Implications for Accounts of the Generation Effect	58
Accounting for Changes in Metamemory Across Blocks.....	60
Limitations and Future Directions.....	61
Conclusion	64
References	65
Appendix 1	71
Appendix 2	73
Appendix 3.....	75

List of Tables

Table 1: Mean proportion of participants who reported changing study strategies across blocks.

Table 2: Strategy reports: Mean proportion of participants using each study strategy for each test.

List of Figures

Figure 1: Experiment 1: Mean proportion of correctly recalled targets on each test for each group. Error bars show the standard error of each mean.

Figure 2: Experiment 1: Mean proportion of correctly recalled targets on each test for each sub-group. Error bars show the standard error of each mean.

Figure 3: Experiment 2A: Mean proportion of correctly recalled read and generate targets across study-test blocks in each group. Error bars show the standard error of each mean.

Figure 4: Experiment 2A: Mean proportion of correctly recalled read and generate targets across study-test blocks for each subgroup group. Error bars show the standard error of each mean.

Figure 5: Experiment 2B: Mean proportion of correctly recalled targets on each test in the R-R/RG-RG group. Error bars show the standard error of each mean.

Figure 6: Experiment 3: Mean proportion of correctly recalled read and generate targets across study-test blocks. Error bars depict the standard error of each mean.

Figure 7: Experiment 3: Mean proportions of correctly recalled read and generate targets across study-test blocks for each subgroup. Error bars represent the error of each mean.

Figure 8: Experiment 3: Proportion of correctly recalled read and generate targets across study-test block in each subgroup. Error bars represent the standard error of each mean.

Figure 9: Experiment 3. Subgroup mean percent of JOLs for read and generate targets across blocks. Error bars depict the standard error of each mean.

Figure 10: Experiment 3: Subgroup mean predictions ratings for predictions for read and generate targets across blocks. Error bars depict standard error of each mean.

Figure 11: Experiment 3: Subgroup mean percent of CJs for read and generate targets across blocks. Error bars depict the standard error of each mean.

Figure 12: Experiment 3: Subgroup mean postdiction ratings for read and generate targets across blocks. Error bars depict the standard error of each mean.

List of Abbreviations and Nomenclature

RG-RG/RG-RG – both read and generate targets are studied and tested in both Blocks

R-R/R-R – read targets are studied and tested in both Blocks

G-G/G-G – generate targets are studied and tested in both Blocks

RG-R/RG-RG – both read and generate targets are studied in both Blocks; Test 1 consists of read targets, Test 2 consists of read and generate targets

RG-G/RG-RG - both read and generate targets are studied in both Blocks; Test 1 consists of generate targets, Test 2 consists of read and generate targets

RG/RG-RG – both read and generate targets are studied in both study-test blocks; no Test 1, Test 2 consists of read and generate targets

R<G – subgroup of the RG-RG/RG-RG group; Test 1 recall greater for generate than read targets

R>G – subgroup of the RG-RG/RG-RG group; Test 1 recall greater for read than generate targets

R=G – subgroup of the RG-RG/RG-RG group; Test 1 recall equivalent for read and generate targets

JOLs – judgments of learning

CJs – confidence judgments

Introduction

People have a multitude of study strategies at their disposal for committing new information to memory, from rote repetition to mnemonic strategies. However, there is a disconnection between how people develop these study strategies, and what we can learn about this development from traditional memory paradigms. Traditional memory paradigms typically include one study experience and one test experience. One study-test experience does not allow researchers to gauge how people learn about the effectiveness of their study strategies and how this learning leads them to modify their future study strategies. My dissertation examined how and when people learn about the effects of a specific study strategy, the *generation effect* (Slamecka & Graf, 1978), and how this knowledge influences a subsequent study-test experience.

Memory research has identified a number of effective encoding strategies including deeper levels of processing, transfer-appropriate processing, production, and generation. The levels-of-processing framework (Craik & Lockhart, 1972) suggests that information studied at a deeper level of processing will be better remembered. For example, Craik and Tulving (1975) showed that participants who encoded words semantically (deep processing) had better memory for the target words when compared to those who encoded words using rhyme sounds (shallow processing). In contrast, the transfer-appropriate processing framework (Morris, Bransford & Franks, 1977) suggests that memory is best when the processing required at test matches how the information was originally studied (e.g., studying fill-in-the-blanks for a fill-in-the-blank test). For example, Fisher and Craik (1977) had people study words either semantically, as categorical pairs, or by rhyming. Those who received a test compatible with how they

studied the words (e.g., semantic encoding then semantic test) showed better memory than those whose test was not compatible with how they studied the words (e.g., semantic encoding then rhyme test). These two frameworks show that both encoding and retrieval processes influence memory outcomes. The levels-of-processing framework emphasizes encoding; in contrast, the transfer-appropriate-processing framework emphasizes the relationship between encoding and retrieval.

The production and generation study strategies require people to be active in the learning process. The production effect suggests that information that is said aloud, as opposed to being read silently, is better remembered (Macleod, Gopie, Hourihan, Neary & Ozubko, 2010). Saying targets aloud makes them distinct relative to reading them silently. The generation effect refers to memory improvements that result from self-generation (e.g., k_t_t_n) relative to simple reading (e.g., kitten; Hirshman & Bjork, 1988; McDaniel, Waddill & Einstein, 1988; Slamecka & Graf, 1978). Generation tasks may involve answering questions, solving word or sentence fragments, or solving anagrams. The generation effect was the encoding manipulation used throughout the current set of experiments (see Bertsch, Pesta, Wiscott & McDaniel, 2007 for a review).

Multiple study-test paradigms expand the utility of memory research to include how experience leads people to adapt their approaches to studying new information. A number of important memorial effects have been identified using multi-trial paradigms including the testing effect (Roediger & Karpicke, 2006), spaced retrieval-practice effects (Whitten & Bjork, 1977), retrieval-induced facilitation (Chan, 2009), and the practice effect (Postman, Burns & Hasher, 1970). Each effect evidences better memory across study-test experiences; however, they occur under very different conditions.

A current topic receiving a lot of attention is the testing effect (e.g., Roediger & Karpicke, 2006), which occurs when study experiences are replaced by additional testing experiences, leading to improved memory on a final test of the same materials. Here, memory is better in a study-test-test paradigm versus a study-study-test paradigm. The spaced retrieval-practice effect (e.g., Whitten & Bjork, 1977) refers to better memory when study-test blocks are spaced rather than massed. Here, memory improves in a study-test/study-test paradigm when the second study-test block occurs after an interval. Retrieval-induced facilitation (e.g., Chan, 2009) also occurs in a study-test-test paradigm. Contrary to the testing effect, improved memory on the final test occurs for previously non-tested materials. More specifically, the items on the first and second test differ; therefore, the act of taking the first test aids performance on the second test. The important similarity among these paradigms is that the same initial study materials are used across all subsequent study and test phases. Additionally, each of these paradigms has shown that testing can also serve as a learning event.

The practice effect (e.g., Postman, Burns, & Hasher, 1970; Schmidt & Bjork, 1992; Snedden, 1931) is most relevant to the current experiment and occurs when memory on a test improves as a consequence of task experience when using a different set of materials across tasks. For example, Postman et al. (1970) showed that recall memory improved across study-test blocks when different study lists were used. Similarly, Gorfein and Blair (1971) found that recall memory improved an average of two items across each of the three study-test trials when different study-lists were used. The effect of task practice (and hence task experience) across study-test blocks is also a focus of the current studies.

Recent research in metamemory has begun to incorporate multiple study-test blocks (e.g., Hertzog, Price & Dunlosky, 2008, Koriat, 1997; Tiede & Leboe, 2009). Metamemory involves higher-order thinking about one's own memory, allowing individuals to monitor cognitive activities requiring the use of memory (Dunlosky & Metcalfe, 2009). For an example of how research in metamemory incorporates a multiple study-test cycle, Koriat (1997) had participants study a series of paired associates (e.g., cow-milk) followed by a cued-recall test. Using different materials, this procedure was repeated. Judgments of learning (JOLs), a measurement of metamemory during study, were also collected during both study tasks, requiring participants to make target-by-target judgments of the likelihood that they would remember the target at test. Both recall accuracy and JOLs increased across tests. Koriat attributed these two increases to a task practice and improved metamemory monitoring, respectively.

Using a multiple study-test design, Dunlosky and Hertzog (2000) examined how experience with study strategies, rather than memory for single targets, influences subsequent memory. Participants studied a series of word pairs using imagery (deep processing) or rote repetition (shallow processing) followed by a cued-recall test. After the first study experience, recall was better for imagery targets than rote-repetition targets, thus words that were studied more *deeply* (Craik and Tulving, 1975) were better remembered on Test 1. Participants were then given another list of word pairs to study using the same strategies, followed by another cued-recall test. The second test experience revealed equivalent recall for both target types, reflecting an increase in recall for words studied using rote-repetition.

Dunlosky and Hertzog (2000) interpreted these findings as evidence of *knowledge updating*; participants became knowledgeable about the relative benefits of imagery during the first task experience and consequently improved their encoding of repetition targets during the second task experience. Importantly, participants were not given any information about the benefits of using imagery as a strategy prior to the experiment, so the mere engagement in the activity was enough for participants to learn about its benefits (see also Brigham & Pressley, 1988). However, Dunlosky and Hertzog noted that their focus was on participants' metamemory judgments (i.e., how participants *think* about their memory), rather than on the study strategies they used.

de Winstanley and Bjork (2004; see also Bjork, de Winstanley & Storm, 2007; Bjork, Storm, & de Winstanley, 2011) used a similar paradigm to investigate whether people (1) are sensitive to the benefits of generation in an initial study-test block, and (2) use this knowledge to adopt a more effective study strategy during a subsequent study-test block. Their paradigm was used and adapted in each of the present experiments. In their Experiment 1A (and 1B), participants studied a paragraph of sentences on a topic, half of which contained a target in red that was *read* silently, and half of which contained a target in red that was presented as a fragment to be *generated*. After a distractor task, participants were then given a fill-in-the-blank test where the same sentences were presented with the critical targets left blank. This Block 1 study-test procedure was then repeated in Block 2 using a different topic.

Critically, a generation effect was found on Test 1 but not on Test 2. The elimination of the generation effect on Test 2 was attributed to improvement for read targets across tests, rather than to impairment for the generate targets across tests. de

Winstanley and Bjork concluded that the elimination of the generation effect on Test 2 was a consequence of the participants learning about the benefits of generation during Test 1, and altering their encoding strategy for the read targets during Study 2 in a manner that led to improved recall of subsequent read targets. Thus, as stated in the title of their article, these participants appear to have become “better readers.”

de Winstanley and Bjork (2004) conducted two follow-up experiments to substantiate their claim that participants must experience the benefits of generation on Test 1 to produce this pattern. Experiment 2 used a between-lists design: participants studied the targets using a single strategy (read or generate) during Block 1 and the alternating strategy during Block 2. A between-lists generation effect occurred on both tests. Experiment 3 used a between-subjects design: participants either read or generated all the targets on both tests. A between-subjects generation effect occurred on both tests. de Winstanley and Bjork contended that the generation effect persisted on Test 2 in both of these cases because participants did not experience the relative benefits of generation versus reading on Test 1.

Bjork and Storm (2011) extended de Winstanley and Bjork’s (2004, Experiment 1) main findings in three important ways. First, the elimination of the generation effect on Test 2 was shown to persist with a 15-minute delay either after Test 1 (Experiment 1) or between each study and test phase (Experiment 2), suggesting the findings from this paradigm could have both applied and educational relevance (e.g., for students studying for exams). Second, they identified a condition under which the generation effect persists on Test 2, namely when participants were given a free recall test for Test 1 (Experiment 3). They argued that a free recall test did not enable participants to learn that connecting

each target word with its sentence (i.e., a *context strategy*) could benefit memory, thus they did not modify their study strategy on Block 2. Third, they provided evidence that participants use this context strategy (Experiment 4). Instead of being tested on the studied targets in Test 2, participants were tested on a sentence context word. Participants given a fill-in-the-blank Test 1 recalled more context words on Test 2 than participants given a free-recall Test 1; they also recalled as many context words for read sentences as for generate sentences on Test 2. Thus, those who were exposed to the sentence contexts during Test 1 (fill-in-the-blank test) were more likely to realize the benefit of context as a strategy when approaching Block 2. Bjork and Storm contend that this increased use of context likely contributes to the elimination of the generation effect on Test 2. However, they did not directly show that increased use of the context strategy is linked to greater target recall. Therefore, it is possible that the increased use of the context strategy increases context recall, but another study strategy shift is responsible for increased recall of read targets on Test 2.

To summarize, de Winstanley and Bjork (2004) posited that the elimination of the generation effect on Test 2 is due to improved recall for read targets. Further, they suggest that taking Test 1 is the triggering factor for participants to improve their study strategy for read targets. Finally, Bjork and Storm (2011) further suggest that participants achieve this end by increasing their use of context as a study strategy for read targets in Block 2. I report three experiments that provide a detailed evaluation of these claims.

Following from de Winstanley and Bjork (2004), participants in each of the present experiments studied a paragraph on a different topic in each of two study-test blocks. Each sentence included a target word highlighted in red that either had to be read

or generated, depending on the group. Memory for the targets was then tested using a fill-in-the blank test. This procedure was repeated in Block 2. In Experiment 1, a within-subject group studied and were tested on both read and generate targets in each block, as in de Winstanley and Bjork (2004, Experiment 1). Two between-subject groups were also tested, receiving only read or generate targets in each block, as in de Winstanley and Bjork (2004, Experiment 3).

The main goal of Experiment 1 was to test whether the increase in recall of read targets across tests was larger in the within group than the read group. This should occur if experiencing generation boosts read target recall. It was also of interest whether the increase in recall of generate targets would be smaller in the within group than the generate group. This might occur if the within group shifts its focus away from the generate targets during Block 2, thus impairing their recall. de Winstanley and Bjork (2004) tested their within and between groups in separate experiments using different stimuli and thus were unable to evaluate these possibilities.

To further evaluate de Winstanley and Bjork's (2004) claim that experiencing the generation effect on Test 1 causes its elimination on Test 2, the within group was also subdivided into three subgroups based on their relative recall of read versus generate targets on Test 1. According to de Winstanley and Bjork's claim, only participants who show the generation effect on Test 1 recall should not show it on Test 2.

Finally, Experiment 1 also evaluated participants' self-reported study strategies. As described above, Bjork and Storm (2011) found some evidence for participants increasing their linking of the targets to their sentence contexts (i.e., *context strategy*). A post-experiment self-report strategy questionnaire allowed me to evaluate what study

strategies people use, whether they shift strategies across blocks, and whether the read targets in particular show a shift in strategy reflecting an increased use of the sentence contexts in the within group.

Experiment 2 further tested de Winstanley and Bjork's (2004) claim that participants become aware of the benefits of generation during Test 1. This surmise was tested by comparing the usual within group to three new groups who were tested only on generate targets or only on read targets on Test 1, or who did not receive a Test 1 at all. If participants learn about the relative benefits of generation during Test 1, then only the usual within group should show the elimination of the generation effect on Test 2. In contrast, if participants learn about the benefits of generation during Study 1, then all groups should show the elimination of the generation effect on Test 2. This latter possibility, though not entertained by de Winstanley and Bjork, is consistent with the finding that metamemory judgments are greater for generate than for read targets during study (e.g., Mazzoni & Nelson 1995; Experiment 2), and is predicted by Dunlosky and Hertzog's (2000) knowledge updating hypothesis.

Experiment 3 collected metamemory judgments at various points in the paradigm (during Study 1, after Study 1, during Test 1, after Test 1; again for Block 2) to determine whether and when participants become sensitive to the relative benefits of generation, and whether shifts in recall across blocks are matched by shifts in metamemory judgments. I was also interested in whether each subgroup's metamemory judgment pattern would match their recall pattern on Test 1. As described in Experiment 3, the metamemory measures used were judgments of learning (*JOLs*) during Study 1, global differentiated predictions (hereafter, *predictions*; Hertzog et al., 2009) after Study 1,

confidence judgments (*CJs*) during Test 1, and global differentiated postdictions (hereafter, *postdictions*; Hertzog et al., 2009) after Test 2.

Previous research has shown metamemory measures to be moderately accurate (e.g., Dunlosky & Metcalfe, 2009, Koriat, 1997). For example, Dunlosky and Nelson (1994) reported JOL sensitivity to targets studied via imagery versus rote repetition, and the pattern of JOLs reflected the pattern of recall (see also Hertzog et al., 2008).

Additionally, Bjork and Storm (2011) have shown people's retrospective metamemorial awareness for the use of context as a study strategy to be moderately predictive of their recall pattern. In the current study, if participants are able to learn about the benefits of generation during the study experience, then this should be reflected in differences in JOLs during Study 1 and/or predictions after Study 1, with higher ratings for generate targets. Or, if participants learn about the benefits of generation at test this should be reflected by greater CJs during Test 1 and/or postdictions after Test 1. Thus, the inclusion of metamemory measures in de Winstanley and Bjork's (2004) paradigm should reveal *when* participants are learning about the generation effect.

Experiment 1

Experiment 1 examined de Winstanley and Bjork's (2004) finding that the generation effect is eliminated after participants experience it on an initial test. The group nomenclature I use identifies the type(s) of targets on the study list and test list in each study-test block (R for *read*, G for *generate*). The *RG-RG/RG-RG group* performed two study-test blocks in which read and generate targets appeared in each study list and on each test. This group was expected to replicate de Winstanley and Bjork's pattern: a generation effect on Test 1, and its elimination on Test 2 due to improved recall of read

targets relative to Test 1, rather than to a decrease in recall for generate targets. I also report detailed subgroup analyses based on participants' relative recall of read versus generate targets on Test 1. According to de Winstanley and Bjork (2004), only the subgroup that shows a generation effect on Test 1 should show the elimination of the generation effect on Test 2.

Experiment 1 also included an *R-R/R-R group*, in which only read targets were studied and tested in each study-test block, to gauge the extent to which recall for read targets improves across tests when the generation effect is not experienced. Experiment 1 also included a *G-G/G-G group*, in which only generate targets were studied and tested in each study-test block. The goal of this group was to examine whether memory for generate targets in the *RG-RG/RG-RG group* suffers a hidden cost across tests. That is, although de Winstanley and Bjork found equivalent recall for generate targets across tests in this group, recall might *increase* across tests in the *G-G/G-G group*. Generate targets in the *RG-RG/RG-RG group* might not show an improvement across tests if participants shift their encoding focus during Study 2 toward the read targets in an attempt to compensate for poorer recall of those targets on Test 1. de Winstanley and Bjork (2004) tested pure-list groups, but they did so in a separate experiment using different materials, and they did not compare this group's recall to their *RG-RG/RG-RG group*.

Finally, half of each group received a questionnaire after the experiment to assess their awareness of the generation effect and the self-reported study strategies used for each study-test block. The questionnaire also allowed me to examine Bjork and Storm's (2011) claim that the elimination of the generation effect on Test 2 is due to an increased linking of targets with their sentence contexts during Study 2.

Method

Participants

University of Calgary undergraduate students (288 in all) received course credit for participating. They were enrolled in at least one psychology course and were drawn from a pool of 1948 potential participants (1387 females, 561 males) whose mean age was 20.4. One participant was replaced for not completing the study task. Participants gave informed consent and were randomly assigned to one of the three groups (96 per group). The recall data were averaged across a replication (48 per group).

Materials

de Winstanley and Bjork's (2004; Experiment 1B) study materials were used, consisting of two paragraphs, one about motivation and goal orientation and one about Bloom's taxonomy of instructional objectives. These materials were also used in Bjork and Storm (2011), and are relevant to students taking psychology. The use of paragraphs also made the paradigm somewhat akin to studying material from a textbook before taking a test. The two paragraphs were not originally matched for difficulty, but assignment of paragraphs to study-test blocks was counterbalanced across participants to control for any materials-based effects.

Each paragraph was divided into 12 sentences presented one at a time during study. The first two sentences in each paragraph were buffers; the last 10 critical sentences contained a target word. The sentences were presented in black and the critical targets were presented in red (see Appendix 1). Each target was studied one of two ways. On *read* trials, participants simply read each target. For example, given the sentence "The emotional or affective part," the target "affective" appeared in red. On *generate* trials, the

target appeared in red as a word fragment for participants to solve (e.g., “The emotional or aff-ct-v- part.”).

The memory test was a fill-in-the-blank test (see Appendix 2), which is a common method of assessing learning of textbook materials. The 10 critical sentences were presented in the order they were studied with a blank space replacing each target word (e.g., “The emotional or _____ part”).

Procedure

Participants were tested in small groups. Participants were told they would study two paragraphs on differing topics, one sentence at a time, and that each paragraph would be followed by an unspecified memory test. They were also told that one word in each sentence (except the first two) would include a word in red, and they would have to try and remember this word. The groups receiving generate trials were told that the target word in red would (sometimes) be presented as a word fragment to be solved, and an example was provided.

During Study 1, the RG-RG/RG-RG group studied a paragraph containing 5 read sentences and 5 generate sentences presented in a set order. Read and generate sentences alternated in an order that was counterbalanced across participants. The R-R/R-R group studied 10 read sentences and the G-G/G-G group studied 10 generate sentences. Each sentence appeared on a projector screen in turn for 15 s. Participants were asked to write down each target (or word-fragment completion) on a separate page of a response booklet. After they wrote down a target, they were instructed to turn the page to prepare to write down the next target. After the 10 sentences were presented participants engaged in a 2-minute distracter task (math problems). Test 1 immediately followed. Participants

were given two minutes to complete the fill-in-the-blank test sheet. The experiment continued immediately with the second study-test block conducted identically with a different paragraph. The entire testing procedure took less than 30 minutes.

Half of each group (48 per group) was also given a post-experiment study strategy questionnaire immediately after Test 2 (see Appendix 3). The questionnaire for the RG-RG/RG-RG group began with a question assessing participants' awareness of the generation effect during Study-Test 1: "What did you notice about how you did on the test of the *first* paragraph?" (after de Winstanley and Bjork; 2004, Experiment 1B). Second, participants retrospectively self-reported their study strategies for read and generate target targets, independently, for each study-test block. Participants were also asked to make an explicit yes/no response to a question probing whether they changed study strategies from Test 1 to Test 2. The R-R/R-R and G-G/G-G groups were given a similar post-experiment questionnaire; however, the questions were worded differently in light of participants having studied only one type of target.

Results and Discussion

Recall scoring followed de Winstanley and Bjork (2004), thus a target was counted as recalled only if it was written within the correct sentence on the test. A lenient method of scoring, where a target was counted as recalled if it was written within any sentence, was also analyzed (see also Bjork & Storm, 2011). The lenient scoring method produced higher recall but did not affect the observed pattern of results and thus is not reported. Effects were significant at $p < .05$ unless otherwise stated.

Three sets of results are presented. First, I evaluated recall in the RG-RG/RG-RG group to establish a replication of de Winstanley and Bjork (2004, Experiment 1). In this

section, recall in the R-R/R-R and G-G/G-G groups was also analyzed and compared to the RG-RG/RG-RG group to gauge whether improvement across tests is linked solely to experiencing the generation effect on Test 1. Second, the RG-RG/RG-RG group was divided into three subgroups based on their Test 1 recall to further evaluate de Winstanley and Bjork's contention that the elimination of the generation effect on Test 2 is a consequence of experiencing the effect on Test 1. Third, the self-reported study strategy data were examined to explore whether different groups/subgroups used different study strategies and in particular whether the RG-RG/RG-RG group was particularly likely to use a context strategy for studying read targets during Block 2.

Recall in the within and between groups

Figure 1 shows the mean correct recall proportions for each group for each study-test block. The RG-RG/RG-RG group's recall was analyzed using a 2 (target type: read, generate) x 2 (test: 1, 2) repeated-measures analysis of variance (ANOVA). The ANOVA revealed a main effect of target type, $F(1, 95) = 19.19$, $MSE = .04$, indicating higher recall for generate than read targets (.48 vs. .38; a generation effect), and a main effect of test, $F(1, 95) = 24.89$, $MSE = .05$, indicating an increase in recall from Test 1 to Test 2 (.37 vs. .49). These main effects were qualified by a significant interaction, $F(1, 95) = 7.72$, $MSE = .06$. Follow-up tests showed a significant generation effect on Test 1 (.48 vs. .29), $t(95) = 4.97$, $SE = .03$, but not on Test 2 (.50 vs. .45), $t(95) = .79$, $SE = .03$, $p = .43$. Additional follow-up tests showed that the increase in recall across tests was significant for read targets (.29 vs. .48), $t(95) = 6.16$, $SE = .03$, but not for generate targets (.45 vs. .50), $t(95) = 1.40$, $SE = .04$, $p = .16$. This pattern fully replicates de Winstanley and Bjork

(2004, Experiment 1), and indicates that the elimination of the generation effect on Test 2 reflected an increase in read target recall, not a decrease in generate target recall.

Figure 1. Experiment 1: Mean proportion of correctly recalled targets on each test for each group. Error bars show the standard error of each mean.

Next, I compared recall of read targets in the RG-RG/RG-RG group (who experienced the generation manipulation) to the R-R/R-R group (who did not). A 2 (group: RG-RG/RG-RG, R-R/R-R) x 2 (test: 1, 2) mixed-factor ANOVA revealed a main effect of group; recall was higher in the R-R/R-R group than in the RG-RG/RG-RG group (.47 vs. .38), $F(1, 190) = 8.42$, $MSE = .09$. Recall increased from Test 1 to Test 2 (.35 vs. .50), $F(1, 190) = 40.61$, $MSE = .05$. Critically, the increase in recall of read targets across tests was similar for both groups thus the interaction was not significant, $F(1, 190) = 2.10$, $MSE = .05$, $p = .15$. Recall in the R-R/R-R group increased across tests (.41 vs. .52), $t(95) = 3.19$, $SE = .04$, even though they did not experience a generation

manipulation. Therefore, the increased recall of read targets across tests in the RG-RG/RG-RG group was not due entirely to experiencing the generation effect.

The equivalent increase in recall in the R-R/R-R and RG-RG/RG-RG groups suggests that both groups shifted study strategies after experiencing the first the fill-in-the-blank test. As discussed below, on this test, both groups may have realized the importance of attending to the sentence contexts, and may have shifted their strategy during Study 2 accordingly, resulting in greater success on their second fill-in-the-blank test. If so, then learning from the generation effect may not be as important in this paradigm as learning from the memory test itself (cf. de Winstanley & Bjork, 2004). Certainly experiencing the generation effect is not the only factor driving increased recall across tests. Changing study strategies as a consequence of experience with a given test is also consistent with Dunlosky and Hertzog's (2000) knowledge updating hypothesis. This interpretation nonetheless has educational implications, such that as students gain experience with a test format across a semester they may become "better test-takers."

Next, recall of generate targets in the RG-RG/RG-RG and G-G/G-G groups was compared to determine if there was a hidden cost to generate targets across tests in the RG-RG/RG-RG group. A 2 (group: RG-RG/RG-RG, G-G/G-G) x 2 (test: 1, 2) mixed-factor ANOVA revealed only a main effect of test: recall increased across tests (.43 vs. .49), $F(1, 190) = 7.90$, $MSE = .05$. Recall was similar in the RG-RG/RG-RG and G-G/G-G groups (.48 vs. .44), $F(1, 190) = 1.41$, $MSE = .08$, $p = .24$. The interaction was not significant, $F < 1$. However, although recall of generate targets did not increase across tests in the RG-RG/RG-RG group (see above), it did increase in the G-G/G-G group (.41

vs. .48), $t(95) = 3.07$, $SE = .02$. Therefore, the elimination of the generation effect on Test 2 in the RG-RG/RG-RG group in part reflected a hidden cost to the generate targets.

Finally, the R-R/R-R and G-G/G-G groups were also compared to determine if a between-subject generation effect occurred. A 2 (group: R-R/R-R, G-G/G-G) x 2 (test: 1, 2) mixed-factor ANOVA revealed an increase in recall across tests (.41 vs. .50), $F(1, 190) = 17.22$, $MSE = .05$. Interestingly, recall was similar in the read and generate groups (.46 vs. .45), $F < 1$, and there was no interaction with test, $F < 1$. Thus, a between-subject generation effect was absent here (cf. de Winstanley & Bjork, 2004; Experiment 3), but this is not unprecedented (e.g., Begg & Snider, 1987; Begg, Vinski, Frankovich, & Holgate, 1991; de Winstanley & Bjork, 1997; de Winstanley, Bjork & Bjork, 1996).

To summarize, the RG-RG/RG-RG group showed improved recall of read targets across tests, but the R-R/R-R group who did not experience the generation effect showed a similar level of improvement. Furthermore, recall of generate targets improved across tests in the G-G/G-G group but not in the RG-RG/RG-RG group, indicating that the elimination of the generation effect on Test 2 in the latter group was in part due to impaired recall of generate targets.

Recall in the within group subgroups

Following Test 1, de Winstanley and Bjork (2004; Experiment 1B) asked participants “What did you notice about your performance on the previous memory test?” The *aware* subgroup (17/31) reported awareness of the generation effect on Test 1, and 13 of these 17 (76%) did not have higher recall for generate (than read) targets on Test 2. The *unaware* subgroup (14/31) did not report awareness of the generation effect on Test 1, and 8 of these 14 (57%) did not have higher recall for generate (than read) targets on

Test 2. Although de Winstanley and Bjork (2004) took these data as support for their account, this difference (76% vs. 57%) is not especially strong evidence that elimination of the generation effect on Test 2 was due to awareness of the effect on Test 1, particularly given the small sample size.

I posed a similar question in my post-experiment questionnaire (e.g., “What did you notice about how you did on the test on the *first* paragraph?”), but only 1 participant (2.1%) reported awareness of the generation effect on Test 1. This low rate is likely due to my posing the question after Test 2 rather than immediately after Test 1 (cf. de Winstanley and Bjork, 2004), which may have led to confusion about the question and/or retrospective memory errors. I therefore adopted a different approach to evaluating whether experiencing the generation effect leads to a strategy shift that selectively improves recall for read targets. Rather than comparing recall based on reported awareness of the generation effect, I used objective Test 1 recall rates for each target type to divide the RG-RG/RG-RG group into three subgroups (see Figure 2). The $R < G$ subgroup ($n = 54$) recalled more generate targets than read targets on Test 1 (i.e., generation effect), the $R > G$ subgroup ($n = 20$) recalled more read targets than generate targets on Test 1 (i.e., negative generation effect), and the $R = G$ subgroup ($n = 22$) recalled an equal number of read and generate targets on Test 1.

Although the generation effect is a robust finding (see Bertsch et al., 2007), a sizeable $R > G$ subgroup was not unexpected. Generation typically induces item-specific processing (focusing attention on the target word alone) whereas reading typically induces relational processing (taking into consideration the meanings of the sentences and the relations between targets and sentences (e.g., McDaniel, Waddill, & Einstein,

1988). To be successful on the fill-in-the-blank test, participants must be able to recall the target word given the sentence cue, thus this test likely benefits from relational processing which would be lacking for generate targets. Therefore, participants who relied on item-specific processing for generate items could show poorer recall for generate than read targets.

Test 2 recall in these subgroups was compared using a 3 (subgroup: $R > G$, $R < G$, $R = G$) $\times$ 2 (target type: read, generate) mixed-factor ANOVA. The key new test here involves the subgroup factor. Test 2 recall was not reliably different across the $R < G$, $R > G$, and $G = R$ subgroups (.46 vs. .53 vs. .53), $F(2, 93) = 1.38$, $MSE = .09$, $p = .26$, nor was the key interaction with target type significant, $F(2, 93) = 1.30$, $MSE = .05$, $p = .28$. Recall for read and generate on Test 2 was similar in the $R < G$ subgroup (.47 vs. .45), $t < 1$, the $R > G$ subgroup (.50 vs. .56), $t < 1$, and in the $R = G$ subgroup (.48 vs. .58), $t(21) = 1.35$, $SE = .07$, $p = .19$. Thus, there was no reliable evidence that the pattern of recall on Test 1 influenced whether a generation effect occurred on Test 2. This result runs contrary to de Winstanley and Bjork's (2004) claim that only participants who experience the generation effect on Test 1 should show the elimination of this effect on Test 2.

Next, I examined how recall within each subgroup shifted across tests using separate 2 (target type: read, generate) $\times$ 2 (test: 1, 2) repeated-measures ANOVAs. For the $R < G$ subgroup, recall increased across tests (.39 vs. .46), $F(1, 53) = 5.58$, $MSE = .04$, and was qualified by a significant interaction with target type, $F(1, 53) = 68.14$, $MSE = .04$. Recall of read targets increased across tests (.20 vs. .47), $t(53) = 7.36$, $SE = .03$, whereas recall of generate targets decreased across tests (.58 vs. .47), $t(53) = 14.35$, $SE = .03$. Experiencing the generation effect on Test 1 made "better readers," replicating the

aware subgroup reported by de Winstanley and Bjork (2004); however, it also made this subgroup “worse generators” on Test 2, which de Winstanley and Bjork, nor Bjork and Storm (2011) did not find. Here, generate targets suffered a cost as a consequence of the improvement for read targets.

Figure 2. Experiment 1: Mean proportion of correctly recalled targets on each test for each sub-group. Error bars show the standard error of each mean.

In the R>G subgroup, recall also increased across tests (.35 vs. .53), $F(1, 19) = 17.44$, $MSE = .04$, and was qualified by a significant interaction, $F(1, 19) = 17.09$, $MSE = .03$. Here, recall of read targets did not change (.48 vs. .50), $t < 1$, whereas recall of generate targets increased across tests (.21 vs. .56), $t(19) = 7.00$, $SE = .05$. Experiencing a negative generation effect on Test 1 made “better generators” but did not make “worse readers” on Test 2. Importantly, these two subgroups compensated for whichever type of target was poorly recalled on Test 1. It seems that experiencing the generation effect is

not the only Test 1 outcome that spurs improved recall across tests; the same is true of experiencing a negative generation effect.

Recall also increased across tests in the R=G group (.35. vs. .53), $F(1, 21) = 8.89$, $MSE = .09$; however, this time the interaction with target type was not significant, $F(1, 21) = 1.82$, $MSE = .03$, $p = .19$. Recall increased across tests both for read targets (.35 vs. .48), $t(21) = 2.19$, $SE = .06$ and for generate targets (.35 vs. .58), $t(21) = 2.89$, $SE = .08$. Participants who experienced equivalent recall on Test 1 became “better readers” and “better generators” on Test 2.

For each subgroup, I also compared the change in recall across tests for a given target type to the R-R/R-R or G-G/G-G group. Thus, the new test here is whether each subgroup’s change in recall across tests for a given target type differed from the change in the relevant between group (i.e., the interaction of group and test). For the R<G subgroup, the increase in read target recall across tests was greater than in the R-R/R-R group (.27 vs. .12), $F(1, 148) = 7.20$, $MSE = .08$. Thus, recall of read targets *can* benefit from experiencing the generation effect relative to the improvement shown in the R-R/R-R group who did not experience the effect, in line with de Winstanley and Bjork (2004). Recall of generate targets decreased across tests in the R<G subgroup, whereas it increased across tests in the G-G/G-G group, resulting in an interaction (-.13 vs. .07), $F(1, 148) = 23.64$, $MSE = .03$. This interaction provides additional evidence that experiencing the generation effect on Test 1 benefits Test 2 recall for read targets, but causes a deficit in Test 2 recall for generate targets (i.e., a trade-off pattern).

The R>G subgroup showed less improvement across tests for read targets than the R-R/R-R group (.02 vs. .12), but this interaction did not reach significance $F(1, 114) =$

1.26, $MSE = .06$, $p = .26$. On the other hand, the increase in recall across tests for generate targets was far greater in the R>G subgroup than in the G-G/G-G group (.35 vs. .07), $F(1, 114) = 25.32$, $MSE = .03$. Thus, experiencing a negative generation effect on Test 1 consequently led to a large improvement in recall for generate targets on Test 2, relative to the small increase shown by the G-G/G-G group.

Finally, in the R=G subgroup, the increase in recall for read targets across tests was similar to the increase in the R-R/R-R group (.13 vs. .12), $F < 1$, but the increase for generate targets was greater than in the G-G/G-G group (.23 vs. .07), $F(1, 114) = 7.16$, $MSE = .03$. This subgroup improved their recall for both target types on Test 2, but this increase was larger than the increase in the between group only for generate targets.

To summarize, none of the subgroups within the RG-RG/RG-RG group showed a generation effect on Test 2, contradicting de Winstanley and Bjork's (2004) claim that elimination of the effect on Test 2 is due to experiencing it on Test 1. Further, the subgroup analyses also showed that the recall pattern on Test 1 determined which target type(s) benefited (or suffered) across tests. Participants who experienced the generation effect on Test 1 (R<G subgroup) improved their memory for read targets (and showed a cost to generate targets) on Test 2; conversely, participants who experienced a negative generation effect on Test 1 (R>G subgroup) improved their memory for generate targets. Importantly, these recall improvements were larger than for participants who only studied one type of target (between-subjects groups). Finally, experiencing equal memory recall for read and generate targets on Test 1 led to parallel memory improvement for both target types on Test 2, such that they showed a generation effect on neither test.

Self-reported study strategies

Using a self-report post-experiment questionnaire, I assessed whether participants felt they had made a study strategy shift across tests and I also asked them to describe their self-reported study strategy for each target type on each test. These reports allowed me to: (1) identify whether participants report changing study strategies across tests, (2) identify the study strategies participants report using, (3) evaluate whether the RG-RG/RG-RG group was particularly likely to shift to a strategy of linking read targets with their sentences during Study 2 (i.e., *context strategy*, as suggested by Bjork & Storm, 2011), and (4) identify whether other types of study strategies increase or decrease in use across study-test blocks.

As Table 1 shows, the majority of participants in each group/subgroup reported changing their study strategy from Study 1 to Study 2. Most participants actively sought to improve their recall across tests, thus improved recall on Test 2 was likely due to more than a passive benefit of task practice. Importantly, the R-R/R-R and G-G/G-G groups also report changing study strategies across blocks, thus changing study strategies was not limited to participants exposed to the generation manipulation.

Given that the study-strategy questions were open-ended, a coding scheme was developed based on categorical groupings of participants' statements (see Delaney & Knowles, 2005; Sahakyan & Delaney, 2003). For both read and generate targets on each test, similar study strategies were grouped together to form a category. Using this method, five study strategy categories emerged: (1) use of sentence context (*context*; e.g., "I tried to remember the whole sentence"), (2) thinking about the target's meaning and/or creating a mental image of the target's referent (*meaning*; e.g., "I thought of the meaning of the red word"), (3) silently repeating the target (*repetition*; e.g., "I repeated the word

over and over in my head”), (4) focusing only on the target (*target only*; e.g., “I just focused on writing down the red word”), and (5) no strategy reported/claimed no strategy used (*no strategy*; e.g., “I didn’t use any particular strategy”).

Table 1

Mean proportion of participants who reported changing study strategies across blocks.

Group/Subgroup	Reported Strategy Change
RG-RG/RG-RG Group	.54
R<G Subgroup (n = 29)	.55
R>G Subgroup (n= 9)	.44
R=G Subgroup (n = 10)	.60
R-R/R-R Group	.81
G-G/G-G Group	.81

The breakdown of each group/subgroup’s reported strategy use on each test is presented in Table 2. In general, many participants claimed not to use a particular study strategy, and those who did tended to use the “shallower” target-only and repetition strategies rather than the “deeper” context and meaning/imagery strategies (Craik & Tulving, 1975). Most participants reported different strategies for studying for the two tests, although their reports were not always consistent with their yes/no strategy shift question responses. Given the small numbers of participants who reported each strategy, the evaluation of study strategies here is purely descriptive.

Bjork and Storm (2011) suggested that the RG-RG/RG-RG group's increased recall of read targets across tests was due to increased linking of targets with their sentence contexts during Study 2. They measured people's ability to recall context words on Test 2, and they also reported metamemory data consistent with their hypothesis. Most participants given the fill-in-the-blank Test 1 explicitly reported switching to the *context strategy* on Test 2, whereas this rarely occurred among participants given the free recall Test 1. Thus, many participants were *explicitly* aware of using the context strategy, though they may also have increased use of other strategies across blocks.

The strategy data were consistent with Bjork and Storm (2011) only in that reported use of a context strategy for studying read targets increased from Block 1 to 2 (.14 vs. .30) in the RG/RG-RG/RG group. However, use of this strategy also increased across blocks for generate targets in this group (.16 vs. .39), even though recall of generate targets did not improve across blocks. Additionally, there was also an increased use of a repetition strategy across blocks in this group, both for read targets (.27 vs. .39) and generate targets (.22 vs. .34). The parallel increase in use of two different strategies, and for both target types, makes it difficult to claim that a context strategy shift was responsible for the Test 2 improvement for read targets in the RG/RG-RG/RG group. In addition, context strategy use also increased across blocks in the R-R/R-R group (.27 vs. .33), and particularly in the G-G/G-G group (.21 vs. .44), suggesting that all groups paid more attention to context after learning about the importance of the sentences for completing a fill-in-the-blank test.

This consistency suggests that learning about test demands may be more important than learning about a particular study strategy during a test. Regardless of

relative Test 1 recall, all groups and subgroups reported an increase in the use of context on Block 2. Consistent with this idea, Bjork and Storm (2011) reported that those who were given a free-recall Test 1 were less likely to incorporate a context strategy in Block 2 versus those who were given a fill-in-the-blank Test 1. A free-recall test does not benefit from the relational processing that a context strategy would provide, but a fill-in-the-blank test would benefit from such additional relational processing. The increase in the use of a context strategy across all groups in Experiment 1 may be reflective of learning about the task, deemphasizing the importance of learning about their relative Test 1 recall. Thus, students should be encouraged to become better “test-takers” meeting the demands of a task, rather than focusing on a particular study strategy.

In general, use of a context, meaning, or repetition strategy increased across blocks for both target types in all three groups. In contrast, use of a target-only strategy, or not using a strategy, typically decreased across blocks. Because of this variety of strategy shifts, changes in recall cannot be neatly attributed to a single strategy shift. It appears that participants achieved their desired ends (improved recall) through a variety of means (strategy shifts), and may have become better “test takers” rather than focusing on a particular study strategy. In the General Discussion I consider ways that future research might better pinpoint the influence of strategy type on recall in this paradigm.

Table 2

Strategy reports: Mean proportion of participants using each study strategy for each test.

Group/Strategy	Read		Generate	
	Test 1	Test 2	Test 1	Test 2

RG/RG-RG/RG Group

Context	.14	.30	.16	.39
Meaning	.07	.11	.00	.07
Repetition	.27	.39	.22	.34
Target Only	.27	.11	.38	.09
No Strategy	.25	.09	.24	.11

R<G Subgroup ($n = 29$)

Context	.14	.21	.17	.35
Meaning	.06	.03	.00	.00
Repetition	.24	.45	.17	.38
Target Only	.24	.10	.35	.03
No Strategy	.31	.21	.31	.24

R>G Subgroup ($n = 9$)

Context	.22	.11	.22	.33
Meaning	.11	.33	.00	.22
Repetition	.22	.11	.22	.22
Target Only	.11	.11	.22	.11
No Strategy	.45	.33	.33	.11

R=G Subgroup ($n = 10$)

Context	.00	.40	.00	.40
Meaning	.00	.10	.00	.10
Repetition	.30	.30	.30	.20
Target Only	.40	.10	.50	.20
No Strategy	.30	.10	.20	.10

R/R-R/R & G-G/G-G Groups

Context	.27	.33	.21	.44
Meaning	.15	.17	.08	.06
Repetition	.17	.23	.17	.25
Target Only	.35	.27	.31	.21
No Strategy	.06	.02	.23	.04

Experiment 2A

Experiment 2A further examined de Winstanley and Bjork (2004) and Bjork and Storm's (2011) claim that participants learn about the memorial benefits of generation by experiencing differential recall of read and generate targets *during the memory test* (i.e., on Test 1). Although these researchers argued that a test experience in which this learning can occur is necessary, participants could potentially realize *during the study trials* that they are encoding the generate targets more effectively than the read targets. If so, the test experience would serve only to confirm their hypothesis. Alternatively, participants may only need to be tested on one or the other type of target to confirm their hypothesis, even though they would therefore not directly experience the generation effect on Test 1.

To test these possibilities, a replication of the *RG-RG/RG-RG group* was compared to three new groups. The *RG-R/RG-RG group* was tested only on read targets on Test 1, the *RG-G/RG-RG group* was only tested on generate targets on Test 1, and the *RG/RG-RG group* did not receive Test 1. None of these groups directly experienced the relative benefits of generation over reading on Test 1, hence the generation effect should be maintained on Test 2 for each group. In contrast, if participants learn about the benefits of generation during Study 1, then the generation effect should be eliminated on Test 2 for each group.

Method

Participants

An additional 192 University of Calgary undergraduates received course credit for participating. The participants were drawn from RPS participant pool and were randomly assigned to one of the 4 groups (48 per group).

Materials and Procedure

The materials and procedure were the same as in Experiment 1, with the exception of the Test 1 experience for the three new groups. On Test 1, the RG-RG/RG-RG group was, as in Experiment 1, tested on the 5 read and 5 generate targets. The RG-R/RG-RG group was tested only on the 5 read targets, and the 5 sentences with generate targets did not appear on the test. The RG-G/RG-G group was tested only on the 5 generate targets, and the 5 read sentences were removed. Finally, the RG/RG-RG group was not given a Test 1. Instead, they were given a 2-minute Sudoku distractor task in addition to the distractor task already in place between study-test blocks.

Results and Discussion

Four sets of analyses are reported. First, the RG-RG/RG-RG group was examined for a replication of de Winstanley and Bjork (2004) and Experiment 1. Second, the RG-RG/RG-RG subgroups, based on Test 1 recall, were analyzed as in Experiment 1. Third, Test 2 recall was compared across the groups to determine the necessary conditions for eliminating the generation effect on Test 2. Finally, recall across tests was compared for the target type(s) present for each group.

Recall in the within group

The RG-RG/RG-RG group results perfectly matched Experiment 1 and de Winstanley and Bjork (2004; see Figure 3). Across tests there was a generation effect (.43 vs. .35), $F(1, 47) = 8.22$, $MSE = .04$, and recall increased from Test 1 to Test 2 (.35 vs. .44), $F(1, 47) = 7.15$, $MSE = .06$. These main effects were qualified by a significant interaction, $F(1, 47) = 7.86$, $MSE = .04$. Follow-up tests showed a significant generation effect on Test 1 (.43 vs. .27), $t(47) = 4.06$, $SE = .04$, but not on Test 2 (.44 vs. .44), $t < 1$.

Recall increased across tests for read targets (.27 vs. .44), $t(47) = 4.05$, $SE = .04$, but not for generate targets (.43 vs. .44), $t < 1$.

Figure 3. Experiment 2A: Mean proportion of correctly recalled read and generate targets across study-test blocks in each group. Error bars show the standard error of each mean.

Recall in the within group subgroups

As in Experiment 1, the RG-RG/RG-RG group was divided into three subgroups based on Test 1 recall (see Figure 4), $R < G$ ($n = 28$), $R > G$ ($n = 7$), and $R = G$ ($n = 13$), and Test 2 recall was analyzed as in Experiment 1. The ANOVA on Test 2 recall did not yield a significant main effect of subgroup (.50 vs. .41 vs. .34), $F(2, 45) = 2.34$, $MSE = .10$, $p = .11$, nor was the interaction of subgroup with target type significant, $F(2, 45) = 1.93$, $MSE = .04$, $p = .16$. Recall of read and generate targets was similar in the $R < G$ subgroup (.48 vs. .51), $t < 1$, and in the $R > G$ subgroup (.37 vs. .46), $t < 1$, and was marginally higher for read targets in the $R = G$ subgroup (.40 vs. .28), $t(12) = 1.86$, $SE = .07$, $p = .09$. Thus, the generation effect was absent on Test 2 regardless of Test 1 outcome, as in Experiment 1. These main effects were qualified by a significant interaction, $F(1, 47) = 7.86$, $MSE =$

.04. Follow-up tests showed a significant generation effect on Test 1 (.43 vs. .27), $t(47) = 4.06$, $SE = .04$, but not on Test 2 (.44 vs. .44), $t < 1$. Recall increased across tests for read targets (.27 vs. .44), $t(47) = 4.05$, $SE = .04$, but not for generate targets (.43 vs. .44), $t < 1$.

I also examined how recall changed across tests for each subgroup, as in Experiment 1 (see Figure 6). In the $R < G$ subgroup, recall increased across tests (.36 vs. .50), $F(1, 27) = 8.47$, $MSE = .06$, and was qualified by a significant interaction, $F(1, 27) = 27.92$, $MSE = .03$. Recall of read targets increased across tests (.19 vs. .48), $t(27) = 14.37$, $SE = .02$, but recall of generate targets did not (.54 vs. .51), $t < 1$. As in Experiment 1, this subgroup became better readers, but unlike in Experiment 1 they did not become worse generators.

The $R > G$ and $R = G$ subgroups should be interpreted with caution given the small subgroup sizes. Recall in the $R > G$ subgroup did not increase across tests (.36 vs. .41), $F < 1$, but the interaction was significant, $F(1, 6) = 7.00$, $MSE = .04$. The decreased recall of read targets across tests was not reliable (.51 vs. .37), $t(6) = 1.70$, $SE = .08$, $p = .14$, nor was the increased recall of generate targets (.20 vs. .46), $t(6) = 1.72$, $SE = .10$, $p = .14$. The interaction suggests that experiencing a negative generation effect on Test 1 lead participants to focus more on the generate targets, and consequently to incur a cost to read targets on Test 2, replicating Experiment 1. Recall in the $R = G$ subgroup did not increase across tests (.31 vs. .34), $F < 1$, and the interaction with target type was marginal, $F(1, 12) = 3.46$, $MSE = .01$, $p = .09$. Recall increased nonsignificantly across tests for read targets (.31 vs. .40), $t(12) = 1.25$, $SE = .07$, $p = .24$, and changed little across tests for generate targets (.31 vs. .28), $t < 1$.

Figure 4. Experiment 2A: Mean proportion of correctly recalled read and generate targets across study-test blocks for each subgroup group. Error bars show the standard error of each mean.

As in Experiment 1, the pattern of recall on Test 1 influenced how recall for each target type shifted on Test 2. Strikingly, as in Experiment 1 the generation effect was eliminated on Test 2 in each subgroup, contrary to de Winstanley and Bjork's (2004) claim that the elimination occurs only when the generation effect occurs on Test 1.

Test 2 recall as a function of the Test 1 experience

If participants must experience the generation effect during Test 1 to modify their encoding strategies in a way that leads to the elimination of this effect on Test 2, then a generation effect should occur on Test 2 in the RG/RG-RG, RG-G/RG-RG, and RG-R/RG-RG groups. Contrary to that possibility, recall of read and generate targets on Test 2 was similar in both the RG/RG-RG group (.32 vs. .36), $t < 1$, and the RG-G/RG-RG group (.46 vs. .50), $t < 1$. However, out of step with this pattern, the RG-R/RG-RG group showed a generation effect on Test 2 (.53 vs. .42), $t(47) = 2.68$, $SE = .04$. Experiment 2B

examines whether the survival of the generation effect on Test 2 following a Test 1 that consists only of read targets is a replicable finding.

To further assess the ostensible benefits of experiencing the generation effect during Test 1, I compared the change in recall across tests in the RG-RG/RG-RG group relative to the RG-R/RG-RG group (for read targets) and to the RG-G/RG-RG group (for generate targets). In each case the key effect is the group by test interaction. The increase in recall of read targets across tests was similar in the RG-RG/RG-RG and RG-R/RG-RG groups (.13 vs. .17), $F < 1$, as was the increase in recall of generate targets in the RG-RG/RG-RG and RG-G/RG-G groups (.01 vs. .02), $F < 1$. Thus, recall of read (or generate) targets did not improve more across tests when both target types were experienced on Test 1 than when only one target type was experienced on Test 1.

Finally, recall of read targets on Test 2 was not greater in the RG-RG/RG-RG group than in the RG-R/RG-RG group (.44 vs. .42), $t < 1$, nor was recall of generate targets on Test 2 reliably lower (.44 vs. .50), $t(94) = 1.09$, $SE = .05$, $p = .28$. Thus, the persistence of the generation effect on Test 2 in the RG-R/RG-RG group was not driven by worse recall of read targets or better recall of generate targets on Test 2 relative to the RG-RG/RG-RG group.

The main finding from Experiment 2A is that the generation effect can be eliminated on Test 2 even when it was not experienced on Test 1 (RG/RG-RG and RG-G/RG-RG groups). This finding challenges de Winstanley and Bjork's (2004) claim that the experience of the generation effect during Test 1 is the seed of its elimination on Test 2. However, a generation effect did occur on Test 2 when only read targets were tested on Test 1 (RG-R/RG-RG group). This finding is discussed further following Experiment 2B.

Finally, recall of read targets improved across tests similarly whether the generation effect was experienced on Test 1 (RG-RG/RG-RG) or not (RG-R/RG-RG group).

Experiment 2B

Experiment 2B followed up the finding that of the four groups tested in Experiment 2A, only the RG-R/RG-RG group showed a generation effect on Test 2. To investigate this exception, while also testing its generality, an *R-R/RG-RG* group was tested. This group studied and were tested on 10 read targets in Block 1, then studied and were tested on 5 read and 5 generate targets in Block 2. If being tested solely on read targets on Test 1 somehow “protects” the generation effect on Test 2, then that pattern should also occur in Experiment 2B.

Method

Participants

An additional 48 University of Calgary undergraduates from the RPS participated for course credit.

Materials and Procedure

The materials and procedure followed Experiments 1 and 2. Block 1 was conducted as for the R-R/R-R group in Experiment 1, and Block 2 was conducted as for the RG-RG/RG-RG group in Experiment 1 and the four groups in Experiment 2A.

Results and Discussion

As can be seen in Figure 5, a generation effect on Test 2 occurred in the R-R/RG-RG group (.47 vs. .56), $t(47) = 1.98$, $SE = .04$, as for the RG-R/RG-RG group in Experiment 2A. Recall of read targets did not increase significantly across tests (.42 vs. .47), $t(47) = 1.29$, $SE = .04$, $p = .20$, unlike for the RG-R/RG-RG group in Experiment

2A. Thus, when participants are tested exclusively on read targets on Test 1—whether they studied both read and generate targets (Experiment 2A) or only read targets (Experiment 2B)—the generation effect appears to be intact on Test 2. Though hard to explain, this pattern is important in showing that the paradigm is sensitive enough to reveal a generation effect on Test 2 under some conditions. Bjork and Storm (2011, Experiment 3) also reported a generation effect on Test 2 when Test 1 was free recall. Thus, at least two conditions lead the generation effect to persist on Test 2.

Test 2 recall in the R-R/RG-RG group was compared to the RG-R/RG-RG group in Experiment 2A via a 2 (group: R-R/RG-RG, RG-R/RG-RG) x 2 (target type: read, generate) mixed-factor ANOVA. The key test here involves the group factor. The difference in recall between the RG-R/RG-RG and R-R/RG-RG groups was not significant (.51 vs. .57), $F < 1$, nor was the interaction with target type, $F < 1$. Thus, the size of the generation effect on Test 2 was similar for the two groups.

Figure 5. Experiment 2B: Mean proportion of correctly recalled targets on each test in the R-R/RG-RG group. Error bars show the standard error of each mean.

The persistence of the generation effect on Test 2 in the RG-R/RG-RG group in Experiment 2A, despite increased recall of read targets across tests, suggests that the encoding of generate targets also improved across tests (or that encoding of generate targets was impaired during Study 2 in the RG-RG/RG-RG group in Experiment 2A). The persistence of the generation effect on Test 2 in the R-R/GR-GR group in Experiment 2B may be a consequence of the introduction of the generation task in Study 2. This group may have been drawn toward encoding the novel generate targets during Study 2. Two findings are consistent with this possibility. First, their recall of read targets did not increase across tests (reported above). Second, their recall of generate targets on Test 2 was greater than in the RG-RG/RG-RG group in Experiment 2A (.53 vs. .43), $t(94) = 2.02$, $SE = .05$, and was equal to the RG-R/RG-RG group in Experiment 2A (.55 vs. .53), $t < 1$. The latter group's attention might also have been drawn toward the generate targets during Study 2 given that target type was not tested on Test 1.

Experiment 3

Experiment 3 examined participants' metamemorial awareness of the generation effect at various stages of the study-test/study-test procedure. Four metamemory measures were collected in a new RG-RG/RG-RG group (Hertzog et al., 2008; 2009) to test de Winstanley and Bjork's (2004) suggestion that participants experience the relative benefits of generation during Test 1, and then use that experience to modify (and improve) their encoding of read targets during Study 2. Analysis of these measures was focused at the subgroup level, given that the different outcomes on Test 1 for each subgroup should be reflected in different metamemory patterns.

First, participants provided *JOLs* during Study 1 to gauge their sensitivity to the relative benefits of generation versus reading. If the R<G subgroup, for example, is sensitive to the benefits of generation prior to Test 1, then JOLs for generate targets should be higher than JOLs for read targets. Second, participants were also asked to make *predictions* regarding their recall of each target type after Study 1 but before Test 1. The predictions measure required participants to predict their recall for each target type (Hertzog et al., 2009) on the upcoming test. If the R<G subgroup, for example, gains knowledge about the benefits of generation versus reading during Study 1 then this should be reflected in higher predictions for generate than read targets.

Third, to evaluate metamemory during Test 1 *CJs* were collected for each recalled target (Hertzog et al., 2009). If JOLs and predictions indicate that the R<G subgroup, for example, is insensitive to the relative benefits of generation versus reading during Study 1, as per de Winstanley and Bjork (2004), then CJs will reveal whether these benefits are noticed *during* Test 1. Fourth, to evaluate the benefits of the Test 1 experience as a whole, participants were also asked to make *postdictions*, that is, predictions about their recall of each target type *after* Test 1 (Hertzog et al., 2009). Postdictions will indicate whether the R<G subgroup, for example, learns of the relative benefits of generation versus reading *after* the completion of Test 1.

If none of the metamemory measures show sensitivity to the Test 1 recall pattern during Block 1, this could suggest that the effects of generation occur implicitly rather than explicitly—a possibility that has not been entertained to date. The same four measures were also administered in Block 2. Given the results of Experiments 1 and 2, it was expected that the generation effect would be eliminated on Test 2 for all three

subgroups ($R < G$, $R > G$, $R = G$), and thus no differences in metamemory judgments on Test 2 were expected. However, I also investigated whether each subgroup's shifts in recall across tests coincided with shifts in their metamemory reports.

Method

Participants

An additional 48 University of Calgary undergraduates from the RPS participant pool participated for course credit.

Materials and Procedure

The materials and procedure were identical to the RG-RG/RG-RG group in Experiment 1 and 2A, with the addition of the metamemory measures. During Study 1 participants provided JOLs after each sentence was studied; they were asked to write down a value on a scale of 0 to 100 indicating their confidence that they would remember that target on an upcoming test (the higher the value, the greater their confidence). After Study 1, but before Test 1, participants provided their predictions by writing down the number of generate targets out of 5, and the number of read targets out of 5, that they thought they would recall. During Test 1 participants made a CJ next to each target they recalled, by assigning a value between 0 and 100 (the higher the CJ, the greater their confidence). CJs were analyzed only for correctly recalled targets. Finally, immediately following Test 1, participants provided their postdictions, by writing down the number of read targets out of 5, and the number of generate targets out of 5 that they thought they correctly recalled. This procedure was then repeated for Block 2.

Results and Discussion

Three sets of analyses are reported. First, the recall data for the RG-RG/RG-RG group was analyzed as in Experiments 1 and 2A. Second, recall in the three subgroups ($R < G$, $R > G$, $R = G$) was evaluated as in Experiments 1 and 2A. Third, each metamemory measure was analyzed at the level of the subgroup to determine if their metamemory ratings mapped onto their recall and onto their shifts in recall across tests.

Recall within the within group

Recall in the RG-RG/RG-RG group was evaluated as in Experiments 1 and 2A. As shown in Figure 6, recall for read and generate targets was similar overall (.35 vs. .36), $F < 1$, and recall increased from Test 1 to Test 2 (.29 vs. .43), $F(1, 47) = 22.07$, $MSE = .04$. The target type by test interaction was in the usual direction but was marginally significant here, $F(1, 47) = 3.00$, $MSE = .05$, $p = .09$.

Figure 6. Experiment 3: Mean proportion of correctly recalled read and generate targets across study-test blocks. Error bars depict the standard error of each mean

Nonetheless, follow-up tests indicated the usual pattern: a generation effect was present on Test 1 (.25 vs. .32), $t(47) = 1.76$, $SE = .04$, $p = .08$, but not present on Test 2 (.45 vs.

.40), $t(47) = 1.08$, $SE = .05$, $p = .29$. And, as in Experiments 1 and 2A, recall of read targets increased significantly across tests (.25 vs. .45), $t(47) = 4.72$, $SE = .04$, but recall of generate targets did not (.32 vs. .40), $t(47) = 1.67$, $SE = .05$, $p = .10$. The addition of the metamemory measures thus did not have much influence on the recall pattern, and, importantly, the generation effect was once again not significant on Test 2.

Recall in the within group subgroups

The RG-RG/RG-RG group was once again divided into three subgroups based on Test 1 recall: $R < G$ ($n = 21$), $R > G$ ($n = 14$), and $R = G$ ($n = 13$). Test 2 was analyzed as in Experiment 1 (refer to Figure 7), and the ANOVA showed that recall of read and generate targets did not differ (.42 vs. .42), $F < 1$. The main effect of subgroup was also not reliable, $F(2, 45) = 2.34$, $MSE = .10$, $p = .11$, nor was the interaction, $F(2, 45) = 1.93$, $MSE = .04$, $p = .16$.

Figure 7. Experiment 3: Mean proportions of correctly recalled read and generate targets across study-test blocks for each subgroup. Error bars represent the error of each mean.

Test 2 recall was nonsignificantly greater for read than generate targets in the $R < G$ subgroup (.46 vs. .36), $t(20) = 1.17$, $SE = .08$, $p = .26$, and the reverse was true in the $R > G$ subgroup (.41 vs. .51), $t(13) = 1.39$, $SE = .07$, $p = .19$. However, Test 2 recall was marginally higher for read than for generate targets in the $R = G$ subgroup (.48 vs. .34), $t(12) = 2.11$, $SE = .07$, $p = .06$. Although these follow-up tests should be interpreted with caution given the small subgroup size, the null interaction here and in Experiments 1 and 2A, do not support de Winstanley and Bjork's (2004) claim that experiencing the generation effect on Test 1 leads to its elimination on Test 2.

Next, the change in recall across tests for read and generate targets was analyzed for each subgroup as per Experiments 1 and 2A. In the $R < G$ subgroup, the ANOVA revealed a significant interaction, $F(1, 20) = 19.79$, $MSE = .04$; the increase in recall of read targets across tests was significant (.12 vs. .46), $t(20) = 5.11$, $SE = .07$, but the decrease in recall of generate targets across tests was not (.44 vs. .36), $t(20) = 1.25$, $SE = .06$, $p = .23$. Participants who experienced a generation effect on Test 1 eliminated the effect on Test 2 by becoming "better readers" (i.e., improving their recall of read targets).

The target type by test interaction in the $R > G$ subgroup was also significant, $F(1, 13) = 17.33$, $MSE = .02$; recall for read targets did not change across tests (.41 vs. .41), $t < 1$, but recall for generate targets sharply improved across tests (.17 vs. .51), $t(13) = 3.62$, $SE = .09$. Thus, participants who experienced a negative generation effect on Test 1 improved recall for generate targets, suggesting they became "better generators."

The interaction in the $R = G$ subgroup was marginally significant, $F(1, 12) = 4.46$, $MSE = .01$, $p = .06$. Recall of read targets increased across tests (.29 vs. .48), $t(12) = 2.65$,

$SE = .07$, whereas recall of generate targets did not (.29 vs. .34), $t < 1$. This subgroup became “better readers” but did not become “better generators.”

Combined Subgroup Analysis

The patterns of recall in the RG-RG/RG-RG subgroups in Experiment 1, 2A, and 3 were very similar but the subgroup results were not identical, perhaps due to the small sample size of some subgroups. In addition, subgroup assignment would be inherently noisy given that there were only 5 targets of each type. Therefore the RG-RG/RG-RG subgroups from Experiments 1, 2A and 3 were pooled here for a combined subgroup analysis (see Figure 8): $R < G$ ($n = 103$), $R > G$ ($n = 41$), and $R = G$ ($n = 48$).

An initial 3 (subgroup: $R < G$, $R > G$, $R = G$) $\times$ 2 (target type: read, generate) $\times$ 3 (Experiment: 1, 2A, 3) mixed-factor ANOVA was conducted on Test 2 recall. Although recall differed across Experiments 1, 2A and 3 (.51 vs. .42 vs. .43), $F(2, 183) = 3.25$, $MSE = .09$, reflecting higher recall in Experiment 1, none of the interactions with Experiment were significant, p 's = .26, .16, and .16. Thus each subgroup is analyzed separately below. Importantly, the difference in recall for read and generate targets on Test 2 was not significant for any subgroup: for the $R < G$ subgroup (.47 vs. .45), $t < 1$, the $R > G$ subgroup (.45 vs. .53), $t(40) = 1.67$, $SE = .05$, $p = .10$, nor the $R = G$ subgroup (.46 vs. .43), $t < 1$. The generation effect on Test 2 in the RG-RG/RG-RG group was absent regardless of the Test 1 recall pattern.

Next, the change in recall across tests in each of the three subgroups was analyzed as described in Experiment 1. In the $R < G$ group, the ANOVA found that overall recall for read targets was less than for generate targets (.32 vs. .49), $F(1, 102) = 105.42$, $MSE = .03$, and increased from Test 1 to Test 2 (.46 vs. .32), $F(1, 102) = 20.96$, $MSE = .05$.

These main effects were qualified by a significant interaction, $F(1, 102) = 114.87$, $MSE = .03$. Recall for read targets increased across tests (.18 vs. .47), $t(102) = 10.76$, $SE = .03$, but decreased for generate targets across tests (.54 vs. .45), $t(102) = 3.17$, $SE = .03$. Those who experienced the generation effect on Test 1 eliminated it on Test 2 by becoming “better readers,” supporting de Winstanley and Bjork’s (2004) claim, but this benefit came with a cost for generate targets (cf. de Winstanley and Bjork, 2004), thus they also became “worse generators.”

Figure 8. Experiment 3: Proportion of correctly recalled read and generate targets across study-test block in each subgroup. Error bars represent the standard error of each mean.

In the R>G group, recall for read targets was greater than for generate targets (.46 vs. .36), $F(1, 40) = 15.80$, $MSE = .02$, and increased from Test 1 to Test 2 (.33 vs. .49), $F(1, 40) = 21.47$, $MSE = .05$. These main effects were qualified by a significant interaction, $F(1, 40) = 42.38$, $MSE = .03$. Across tests, recall for read targets did not change (.46 vs. .45), $F < 1$, but increased for generate targets (.20 vs. .53), $t(40) = 7.11$, SE

= .05. Thus, those who experienced a negative generation on Test 1 eliminated it on Test 2 by becoming “better generators” but not “worse readers.”

Finally, recall for read and generate targets in the R=G group did not differ (.39 vs. .38), $F < 1$, but recall increased from Test 1 to Test 2 (.32 vs. .45), $F(1, 47) = 11.05$, $MSE = .07$. As expected, the interaction was not significant, $F < 1$. The increase was significant for both read (.32 vs. .46) and generate (.32 vs. .43) targets across tests, $t(47) = 3.52$, $SE = .04$, and $t(47) = 2.35$, $SE = .05$, respectively. Those who did not experience a differential recall on Test 1 improved their memory for both types of targets equally across tests; there was no incentive to focus more on one type of target over another.

The combined subgroup analysis confirmed the existence of three distinct types of participants in the within group, each of whom made a unique shift in recall across tests based on their particular Test 1 recall pattern for read and generate targets.

Metamemory measures

Each metamemory measure was analyzed at the subgroup level. First, I examined the metamemory measures collected during Block 1 for read and generate targets, to determine whether and at what point each subgroup become aware of their Test 1 recall pattern. Second, I examined how each subgroup's metamemory measures shifted from Block 1 to Block 2 given their different Test 1 recall outcomes. The Block 2 metamemory measures were not separately analyzed given each subgroup showed similar recall for read and generate targets on Test 2 (hence no differences in metamemory reports were expected).

Metamemory in Block 1. The four metamemory measures are considered in the order they were administered. First, the JOLs during Study 1 (see Figure 9) were

nonsignificantly higher for read than for generate targets for the R<G subgroup (76.1 vs. 75.6), $t < 1$, the R>G subgroup (70.9 vs. 68.1), $t < 1$, and the R=G subgroup (70.0 vs. 66.0), $t(12) = 1.81$, $SE = 2.23$, $p = .10$.

Figure 9. Experiment 3: Subgroup mean percent of JOLs for read and generate targets across blocks. Error bars depict the standard error of each mean.

There was no evidence that the R<G and R>G subgroups were sensitive to the memorial differences between the read and generate targets during Study 1 of their subsequent Test 1 recall pattern.

Second, predictions of recall for read and generate targets collected after Study 1 (see Figure 10) did not differ significantly in the R<G subgroup (3.52 vs. 3.71), $t(20) = 1.28$, $SE = .15$, $p = .21$, the R>G subgroup (3.14 vs. 3.00), $t < 1$, or the R=G subgroup (3.38 vs. 3.15), $t(12) = 1.15$, $SE = .20$, $p = .27$. Thus, participants could not predict a memorial benefit after Study 1 for generate targets (R<G subgroup), or for read targets (R>G subgroup). Taken together, to the extent that these measures were sensitive enough,

participants were not able to predict any memorial differences between read and generate targets of their impending Test 1 recall during or after Study 1.

Figure 10. Experiment 3: Subgroup mean predictions for read and generate targets across blocks. Error bars represent the standard error of each mean.

Third, and in contrast to the first two metamemory measures, CJs collected *during* Test 1 (see Figure 11) were much lower for read targets than for generate targets in the R<G subgroup (32.0 vs. 72.0), $t(20) = 4.41$, $SE = 9.00$, and conversely were much higher for read targets than for generate targets in the R>G subgroup (84.1 vs. 51.1), $t(13) = 2.60$, $SE = 12.70$. The subgroups reported memorial differences during Test 1 of their recall for the two targets types. CJs for read and generate targets were similar in the R=G subgroup (76.2 vs. 80.0), $t < 1$. Thus the CJs in each subgroup paralleled their pattern of Test 1 recall.

These differences in CJs for the two target types do not necessarily evidence *awareness* of the generation effect (R<G subgroup) or the negative generation effect (R>G subgroup), however. Participants were not required to identify whether the target they recalled and rated was a generate or read target, thus we can only confidently say that participants were aware that they had better memory for some targets versus others, not that they had better memory for a given target type over another.

Figure 11. Experiment 3: Subgroup mean percent of CJs for read and generate targets across blocks. Error bars depict the standard error of each mean.

Finally, postdictions obtained *after* Test 1 did not differ significantly for read and generate targets in the R<G subgroup (1.9 vs. 2.0), $t < 1$, the R>G subgroup (2.0 vs. 1.6), $t(13) = 1.71$, $SE = .25$, $p = .11$, or the R=G subgroup (2.0 vs. 1.7), $t(12) = 1.30$, $SE = .24$, $p = .22$ (see Figure 12). That CJs during Test 1 were sensitive to Test 1 recall, but postdictions were not, suggests that the postdiction measure may not have been very sensitive. This possibility is considered in more detail below.

Figure 12. Experiment 3: Subgroup mean postdictions for read and generate targets across blocks. Error bars represent the standard error of each mean

In summary, participants' metamemorial reports in Block 1 suggest they were unable to report (or did not experience) memorial differences between read and generate targets during or after Study 1, or after Test 1, but were able to do so *during* Test 1 (CJs). These findings provide some support for de Winstanley and Bjork's (2004) contention that experiencing the generation effect on Test 1 can be important for modifying subsequent encoding. However, the R>G subgroup also modified their subsequent encoding, thus experiencing a negative generation effect on Test 1 can have a parallel influence. Moreover, Experiment 2A revealed that a Test 1 experience is not necessary for the generation effect to be eliminated on Test 2; this elimination also occurred in the RG-G/RG-RG and RG/RG-RG groups. Even though a difference in CJs was reported, this difference does not imply an awareness of the benefits of generation (or read). Importantly, though, the differences in CJs consistently reflected each subgroup's pattern of recall. The lack of awareness of the generation effect in JOLs, predictions, and

postdictions also suggests that participants may simply be learning about how they should study for a fill-in-the-blank test during Test 1, rather than learning about the effects of generation per se. If so, rather than encouraging students to try to learn about the effectiveness of a given study strategy, educators should encourage students to pay attention to the demands of the test, and thus to tailor their study strategies toward the demands of that test.

Metamemorial changes across blocks. Because the CJs were the only metamemory measure to reliably detect a difference for read and generate targets during Block 1, the shift across Block 1 and Block 2 was only examined for this measure. The specific goal was to determine whether the $R < G$ and $R > G$ subgroups were sensitive to the memorial differences between read and generate targets during Test 2 of the shift in their recall pattern (i.e., to a null generation effect). A 2 (target type: read, generate) $\times$ 2 (test: 1, 2) repeated-measure ANOVA was conducted on CJs for each subgroup. The key test here is the interaction.

In the $R < G$ subgroup, CJs increased from Test 1 to Test 2 (52.2 vs. 73.0), $F(1, 20) = 9.00$, $MSE = 1001.71$, and the interaction was significant, $F(1, 20) = 11.04$, $MSE = 681.83$. CJs increased across tests for read targets (32.4 vs. 72.0), $t(20) = 4.00$, $SE = 10.00$, but not for generate targets (72.0 vs. 73.8), $t < 1$. There was no generation effect on CJs on Test 2 (72.0 vs. 73.8), $t < 1$. Importantly, the increase in CJs for read targets across tests matched the pattern of recall for this subgroup, indicating this subgroup was able to report their shift in recall during Test 2.

In the $R > G$ subgroup, CJs were similar across Test 1 and 2 (67.6 vs. 80.3), $F(1, 13) = 2.34$, $MSE = 1009.34$, $p = .16$, and despite low power the interaction was

marginally significant, $F(1, 13) = 3.82$, $MSE = 859.48$, $p = .07$. CJs for read targets was not reliably different across tests (84.0 vs. 81.4), $t < 1$, but they did increase marginally for generate targets (51.1 vs. 79.1), $t(13) = 1.96$, $SE = 14.29$, $p = .07$. There was no generation effect on CJs on Test 2 (81.4 vs. 79.1), $t < 1$. The R>G subgroup's CJs shifted across tests in parallel with their recall, indicating this subgroup was reliably able to report their shift in recall during Test 2.

In summary, Experiment 3 did not provide reliable evidence that participants were metamemorially aware of their Test 1 recall pattern during Study 1, although the sensitivity of most of the metamemory measures remains in question. However, participants were able to accurately report memorial differences for read and generate targets during Test 1 in their CJs. Even though a difference in CJs does not imply the participants had an explicit awareness of a generation effect (or negative generation effect), the pattern of CJs mapped onto their corresponding patterns of recall across blocks. This suggests that participants were aware of their memory performance and how it changed across blocks. These results support but broaden de Winstanley and Bjork's (2004) contention that the Test 1 experience is an important modifier of future encoding.

General Discussion

de Winstanley and Bjork (2004) reported that once participants experience the benefits of generation on an initial test, they come to realize they need to become "better readers." To this end, participants modify their encoding strategy for read targets in a manner that eliminates the generation effect on a subsequent test. My primary motivation was to evaluate these claims by delimiting how and when the generation effect is eliminated across study-test experiences. To this end, I obtained three successful

replications of de Winstanley and Bjork's (2004) pattern of results (Experiment 1, 2A, 3; see Francis, 2012, for a discussion on the importance of replication). I also showed that the elimination of the generation effect on Test 2 is not an invariant outcome (Experiment 2A/2B; see also Bjork & Storm, 2011, Experiment 3). Thus, there is a replicable and interesting pattern here, and a set of boundary conditions under which it does not occur--both of which will require explanation.

Accounting for the elimination of the generation effect

My experiments indicate that de Winstanley and Bjork's (2004) claims were too restrictive, in five ways: (1) recall improves across tests for between-subject groups who do not experience the generation effect (Experiment 1), (2) experiencing a negative generation effect also leads to shifts in encoding strategies that influence future recall and eliminate the generation effect on Test 2 (Experiments 1-3), (3) elimination of the generation effect on Test 2 in the within group partially reflects impaired recall for generate targets on Test 2 when compared to Test 1 recall (Experiment 3, combined analysis), (4) participants can also learn about their differential recall of read and generate targets during Study 1 rather than having to being tested on both types of targets (Experiment 2A), and (5) the improvement for read targets across blocks cannot be attributed to an increased use of a context study strategy (Experiment 1). Each of these points is discussed in turn.

1. Recall also improves across tests for between-subject groups. In Experiment 1, recall for read targets increased across tests equally for the RG-RG/RG-RG group (who experienced the generation effect) and the R-R/R-R group (who did not). The increase for the R-R/R-R group clearly shows that factors other than generation are at

work in shifting people's strategies across study-test cycles in this paradigm (although this increase could also reflect a task-practice effect). However, the R<G subgroup of the RG-RG/RG-RG group showed greater improvement for read targets across blocks than the R-R/R-R group. This provides some support for de Winstanley and Bjork's (2004) contention that generation can be an effective instigator of strategy shifts. Sensibly, the R>G subgroup of the RG-RG/RG-RG group showed little improvement across tests for read targets, given this group likely focused on improving its recall of generate targets.

With respect to recall of generate targets in Experiment 1, the G-G/G-G group improved across tests, whereas the RG-RG/RG-RG group did not, as discussed below. In addition, recall for generate targets decreased across tests for the R<G subgroup, but increased across tests for the R>G subgroup. Thus, participants in the within group appear to change their processing efforts for *both* read and generate targets across tests.

Given that both the within and between groups show improvement on Test 2 suggests that on Test 1, participants in this paradigm may simply learn about the importance of encoding the sentence contexts for succeeding on a subsequent fill-in-the-blank test. Consistent with this possibility, as discussed below, after taking Test 1, all groups in Experiment 1 were more likely to report using a context strategy. Thus, the present paradigm may be better suited to promoting learning about how to study for a given type of test than to promoting learning about the benefits of a given study strategy such as generation.

2. Experiencing a *negative* generation effect also sponsors study strategy shifts. Experiments 1-3 confirmed that experiencing the generation effect on Test 1 (R<G subgroup) improves future encoding and recall of read targets, but also revealed that

experiencing a *negative* generation effect on Test 1 (R>G subgroup) improves future encoding and recall of generate targets. Most interestingly, these subgroups' opposing shifts both resulted in a nonsignificant generation effect on Test 2. Thus, participants may shift their encoding efforts to whichever target type suffered on Test 1, and both shifts lead to the same Test 2 outcome.

Participants who did not experience differential recall on Test 1 (i.e., the R=G subgroup) also showed an interesting pattern. The combined analysis in Experiment 3 revealed that these participants improved their recall of *both* types of targets on Test 2 (and, as for the other two subgroups, they also did not show a generation effect on Test 2). Having no incentive to shift focus from one target type to the other spurred improvement for both types of targets.

3. Elimination of the generation effect partially reflects a cost to generate target recall. de Winstanley and Bjork (2004) reported that while recall of read targets improved on Test 2, recall of generate targets did not change. Based on this pattern, they argued that the elimination of the generation effect does *not* come at a cost to generate targets on Test 2 (see also Bjork et al., 2007; Bjork et al., 2011). However, the combined subgroup analysis in Experiment 3 showed that those who experienced the generation effect on Test 1 (R<G subgroup) *did* show a decrease in recall across tests for generate targets. Thus, there is evidence for a trade-off in which increased efforts to encode read targets leads to impaired encoding of generate targets. Begg and Snider (1987) showed that read targets are often remembered more poorly than generate targets due to the processing demands of generating. Thus, generating can make “lazy readers” rather than boosting memory. The current results suggest that some participants may have become

“lazy generators” during Block 2 in an effort to improve their memory for read targets. Here the benefits of improving memory for one target type came at a cost to memory for the other target type.

In contrast, the $R > G$ subgroup, who experienced a negative generation effect on Test 1, increased performance for generate targets across tests but with no concomitant impairment for read targets. Thus, there was no trade-off, or shift toward “lazy reading”, when processing efforts were focused on increasing recall for generate targets. This asymmetry may be due to reading being so automatic that memory for these target types is not impaired when processing is diverted toward generating.

4. Learning can also occur at study, not just at test, in this paradigm.

de Winstanley and Bjork (2004) specifically argued that a Test 1 experience is *necessary* in order to eliminate the generation effect on Test 2. In contrast, Experiment 2 showed that participants who did not receive a Test 1, or whose Test 1 included only the generate targets, also showed this Test 2 pattern. Thus, people can also learn about the generation effect during the study phase (although the metamemory measures in Experiment 3 did not reveal awareness of the generation effect during the study phase).

However, learning during at study does not negate the possibility of learning also occurring at test. Indeed, the CJs collected during Test 1 mapped nicely onto each of the subgroups’ recall patterns, evidencing learning during Test 1, consistent with de Winstanley and Bjork’s contention. Additionally, participants in Experiment 2A and 2B who were tested only on read targets on Test 1 showed a generation effect on Test 2, confirming that the Test 1 experience is an important contributor to changes in recall

across blocks. Learning is a chronic mechanism thus it should not be too surprising that it takes place in both the study and test phases in this paradigm.

5. Context strategy use is not a diagnostic predictor of recall shifts across tests. Bjork and Storm (2011, Experiment 4) provided evidence that the improvement in recall for read targets on Test 2 may be due to increased linking of targets with the sentences across blocks (i.e., a context strategy). Participants who were exposed to Test 1 sentence context recalled more context words on Test 2 versus those who were not exposed to Test 1 sentence context. In addition, 58% of participants who received the fill-in-the-blank Test 1 explicitly reported switching strategies to focus more on context during the second block, versus 13% of participants who received a free-recall Test 1. However, a direct link between use of a context strategy and target recall was not made in their study, or in the present study. Thus, the increased use of a context study strategy may lead to increased recall for context words across tests, but not increased recall for target words. It is even possible that different strategy shifts (e.g., repetition) may be responsible for the increased recall of read targets versus context words across tests. In the self-reported study strategies from Experiment 1, I found that participants also increased their use of context across blocks; however, there was no evidence that participants increased their use of context more for read targets than for generate targets across blocks. Additionally, there was also a significant increased use in repetition across blocks. Thus, the within group increased their use of *multiple* study strategies across blocks, so the elimination of the generation effect cannot be attributed to the increased use of context. In addition, because, the between groups who did not experience the

generation effect also increased their use of context, this strategy shift may be ubiquitous, at least when participants realize they are completing fill-in-the-blank tests.

Use of Study Strategies

Experiment 1 added to our understanding of the study strategies used in this paradigm and how those strategies shift across blocks. Participants reported employing a variety of strategies, and perhaps surprisingly, the majority changed the strategy they used across blocks. This was true of both the within and between groups. Fitting with the research on task practice, participants update or change their study strategies after experience with a task (Bjork, 2011; Delaney, Verkoeijen, & Spirgel, 2010; Dunlosky & Hertzog, 2000; Schmidt & Bjork, 1992).

In Experiment 1, use of a context strategy increased across tests. Further, Bjork and Storm (2011, Experiment 4) provided evidence of increased memory for *context* across tests, but did not show directly that greater context use leads to improved recall of *read targets* across tests. Thus, it remains possible that shifting to a context strategy increased context recall, whereas increased target recall may reflect (at least in part) a different strategy shift (e.g., repetition). Additional research is necessary to solidify the link between context use and changes in recall across test experiences.

To this end, participants could be assigned a particular study strategy for each block. For example, one group could be assigned a repetition strategy for Block 1 and a context strategy for Block 2. In addition to evaluating the effectiveness of individual study strategies, the effectiveness of particular strategy shifts could then be evaluated. There were too few participants in Experiment 1 to enable such evaluations. Another approach would be to have participants rate how much they used each of the strategies

(e.g., repetition, context, target only, meaning, or no strategy), and to then use regression to evaluate the extent of which the use of a particular study strategy predict recalls.

Implications for accounts of the generation effect

There are many accounts of the generation effect, including the transfer-appropriate multifactor account (de Winstanley, Bjork, & Bjork, 1996) the procedural account (Crutcher & Healy, 1989), and the distinctiveness account (Hunt, 2003).

Although the focus of my dissertation was on how encoding strategies shift with task experience, and how that influences memory, my results also have implications for these accounts, as discussed in turn next.

The transfer-appropriate multifactor account and the procedural account both emphasize the relationship between the processing used at study and at test. The transfer-appropriate multifactor account (de Winstanley et al., 1996)—an updated version of the multifactor account (McDaniel, Waddill, & Einstein, 1988)—indicates that the benefits of generation emerge when the type of retention test is sensitive to the kind of information that is strengthened by generation at study. On the procedural account, when information is generated (versus read), people are more likely recapitulate or simulate the processing operations they engaged in at study during the test experience (Crutcher & Healy, 1989; McNamara & Healy, 1995).

The generation effect on Test 1 for most of the within group suggests that the fill-in-the-blank test is conducive to recalling generated targets, suggesting there is a good processing overlap between Study 1 and Test 1 for generated targets. Thus, both the transfer-appropriate multifactor (de Winstanley et al., 1996) and procedural (Crutcher & Healy, 1989) accounts would predict a generation effect on Tests 1 and 2. However, in

order for the generation effect to be eliminated on Test 2, these accounts would predict that the encoding of read targets, but not of generate targets, should change during Study 2. A new encoding procedure would have to be applied to the read targets, thus making it more likely that this procedure would be reinstated during Test 2. The current research does not lend support to this possibility given that participants reported changing study strategies for both types of targets across tests. Thus, it is unclear how the transfer-appropriate multifactor and procedural accounts would explain the disappearance of the generation effect on Test 2.

The distinctiveness account (Hunt, 2003; McDaniel & Geraci, 2006) suggests that targets that are encoded uniquely will be better remembered on an upcoming test. Hence this account is consistent with the generation effect. However, to explain the elimination of the generation effect on Test 2, it would have to be posited that the read targets were now encoded more distinctively than generate targets, or that generate targets were encoded less distinctively than read targets. The increased recall for read targets (and the decrease for generate targets) across tests in the within group is consistent with the argument for a distinctive processing shift toward read targets during Study 2. On the other hand, the self-reported study strategies in Experiment 1 suggest that participants change study strategies for *both* read and generate targets across blocks, and it is unclear why or how they would do so in a way that renders the encoding of one type of target less distinctive than the other. And, of course, had they done so those differences in distinctiveness for read versus generate targets should have resulted in either a generation effect or a negative generation effect on Test 2. Thus, it is also difficult to reconcile the findings from this paradigm with the distinctiveness account.

Accounting for changes in metamemory across blocks

Nelson and Naren's (1990) metacognitive model distinguishes between monitoring and control. Monitoring is the metamemorial component, and is captured by metamemory measures as JOLs and CJs (see also Koriat & Goldsmith, 1996). Control is the behavioural component of memory, such as using a study strategy, and is captured by memory tests. On this model, the within group may have learned metamemorially (as represented by their differences in CJs), then adjusted their study strategies to improve their recall. However, this model also predicts that participants should have been able to adjust their behaviour based on monitoring their memory during study (i.e., JOLs), something I did not find.

A potential concern about Experiment 3 is whether the metamemory measures were sensitive enough. Studies using JOLs (e.g., Dunlosky & Metcalfe, 2009) and predictions/postdictions (Hertzog et al., 2009) typically employ 50 or more trials. With only 5 trials per target type in Experiment 3, the possibility that participants can learn about the generation effect (or negative generation effect) during the study phase cannot be ruled out.

According to Koriat (1997), JOLs can be based on intrinsic or extrinsic cues. Intrinsic cues include the characteristics of the target itself, (e.g., meaning) and the perceived ease of remembering it on a subsequent test. Extrinsic cues include the encoding operations (i.e., generate vs. read). Koriat found that people consistently ignore extrinsic factors when making JOLs; thus, it may not be surprising that they did not rate generate targets as more memorable than read targets during study. Therefore, the cue-

utilization approach may help explain why the JOLs were not sensitive to participants' differences in recall in this paradigm.

An alternative possibility is that the Experiment 3 measures (JOLs, predictions, postdictions) were sensitive enough, but participants are simply not explicitly aware of the memorial differences for read and generate targets during or after Study 1. If so, then the shifts in recall across the subgroups may reflect more of an implicit (vs. explicit) form of awareness of their memory performance. Even the differences in CJs for read and generate targets cannot be attributed to an explicit awareness of the memorial differences between the two target types during Test 1. Participants altered their study strategies for Test 2 regardless of their awareness of Test 1 recall. Thus, if explicit awareness of the memorial differences between target types on Test 1 recall is not responsible for recall shifts across tests, then the educational benefits of encouraging students to become aware of their relative Test 1 recall is unclear. On the other hand, learning about the test demands may be more educationally relevant than learning about a particular study strategy. However, previous studies suggest that participants can be aware of the memorial differences between encoding strategies during study. For example, Mazzoni and Nelson (1995; Experiment 2) found that JOLs were significantly higher for generated targets than read targets during study, and importantly, these JOLs reflected actual recall (see also Castel, McCabe, & Roediger, 2007). Replication of Experiment 3 using a larger number of targets (or participants) should help pinpoint when and if metamemorial awareness emerges in this paradigm.

Limitations and Future Directions

Several limitations and future directions have already been identified and suggested (see above). In this section a few others are considered.

First, the sample consisted solely of undergraduates currently taking a psychology class, making the generality of the conclusions unknown. In addition, the pool included both senior-level psychology students as well as those new to psychology courses. Therefore, experience and engagement in psychology could have modulated both the study strategies used and recall. For example, experienced/engaged students may employ better study strategies than those with less psychology/university experience. In addition, some students might have been familiar with the generation effect from a psychology course. Future research could use a more naïve sample or could explore these variables. In addition, psychology students may have knowledge of the topics covered in the two paragraphs used here. Future research using this paradigm could explore how experience with study materials modulates strategy use, metamemory, and recall.

A second limitation concerns the heavy reliance on post-hoc subgroups. Although the subgroups were divided on an objective basis (i.e., Test 1 recall), participants tended to improve across tests for the target type they did worst on. This shift could reflect regression towards the mean rather than a shift in study strategy, even though the subgroups showed consistent, unique shifts across experiments. However, the pattern of CJs mapped onto each subgroup's pattern of recall, suggesting that the subgroups' metamemorial experiences during Test 2 differed across subgroups, which a regression-toward-the-mean explanation would not predict.

However, it must be acknowledged that the subgroups may have differed with respect to other hidden variables other than their Test 1 recall. For example, the $R < G$

subgroup may have been knowledge about the benefits of generation, or may have been more experienced/engaged participants, whereas the R>G subgroup may have included mostly naïve participants who believed that their memory would be worse for generate targets because they provide impoverished perceptual information about targets. Thus, the present study provides no insights into what generated the different Test 1 patterns.

Third, CJs provide no information about whether participants were aware of whether a target was read versus generated at study. Thus, CJ differences for read and generate targets across tests do not provide evidence of awareness of a generation effect (or negative generation effect). Future research could ask participants to make a source-memory judgment for each recalled target along with a CJ, to better link recall with metamemorial awareness. Making both source-judgments and CJs might also lead participants to think about the costs/benefits of using a particular study strategy, thus leading to greater shifts in strategy use and recall across study-test experiences. This experiment would also test whether making people *aware* of memorial differences between read and generate targets (which differences in CJs alone might not evidence) promotes strategy shifts across tests.

Finally, although a key aspect of the present paradigm is that participants learn from their experiences without explicit feedback it would nonetheless be interesting to examine how providing explicit feedback after Test 1 affects Test 2 recall. Feedback after a test has been shown to improve memory on a subsequent test in other paradigms (e.g., Metcalfe & Kornell, 2007). The power of feedback may have an additive effect to experiencing the generation effect. For example, after Test 1, R<G participants informed that they recalled more generate than read targets may receive confirmation of their

intuitions and metamemory monitoring. Such participants might be especially likely to shift focus toward the read targets in Block 2, perhaps to the point of producing a negative generation effect on Test 2. A crossover effect such as this would provide a striking example of how experience can shape subsequent encoding and retrieval. This experiment would again test whether making people aware of their memorial differences for read and generate targets can promote the development of more effective study strategies across tests.

Conclusion

My dissertation provided a detailed examination of how learning about the benefits of one study strategy, generation, affects encoding, recall, and metamemory and subsequent encoding, recall, and metamemory. This research moved beyond measuring the immediate effectiveness of a given study strategy to evaluating how people modify their learning strategies as a function of task experience. This approach to the study of memory could have applicability in educational contexts. After all, the present research shows that people can monitor their own learning from both study and test experiences, and can apply this knowledge adaptively in the future. As learners discover the benefits and costs of a given study strategy, they create a scaffold for developing improved study strategies for themselves and for other learners.

References

- Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 553-563.
- Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generation makes words more memorable, but so does effective reading. *Memory and Cognition*, 19, 487-497.
- Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. (2007). The generation effect: A meta-analytic review. *Memory and Cognition*, 35, 201-210.
- Bjork, R. A. (2011). On the symbiosis of remembering, forgetting and learning. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: A Festschrift in Honour of Robert A. Bjork* (pp. 1-22). New York: Psychology Press.
- Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1113-1124.
- Bjork, E. L., Storm, B. C., & de Winstanley, P. A. (2011). Learning from the consequences of retrieval: Another testing effect. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: A Festschrift in Honour of Robert A. Bjork* (pp. 351-368). New York: Psychology Press.
- Bjork, E. L., de Winstanley, P.A., & Storm, B.C. (2007). Learning how to learn: Can experiencing the outcome of different encoding strategies enhance subsequent encoding? *Psychonomic Bulletin and Review*, 14, 207-211.
- Brigham, M. C., & Pressley, M. (1988). Cognitive monitoring and strategy choice in younger and older adults. *Psychology and Aging*, 3, 249-257.

- Castel, A. D., McCabe, D. P. & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin and Review*, 14, 107-111.
- 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, 153-170.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.
- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104, 268-294.
- Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 669-675.
- de Winstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory and Cognition*, 32, 945-955.
- de Winstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, 4, 31-48.
- de Winstanley, P. A., & Bjork, E. L. (1997). Processing instructions and the generation effect: A test of the multifactor transfer-appropriate processing theory. *Memory*, 4, 31-48.
- Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects

in the free recall of unmixed lists. *Journal of Memory and Language*, 52, 120-130.

Delaney, P. F., Verhoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In. B. H. Ross (Ed.), *The Psychology of Learning and Motivation Vol. 53* (pp. 63-148). Burlington: Academic Press.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging*, 15, 462-474.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Los Angeles: Sage.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, 33, 545-565.

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, 3, 701-711.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, 7, 585-594.

Gorfein, D. S., & Blair, C. (1971). Factors affecting multiple trial free-recall. *Journal of Educational Psychology*, 1, 17-24.

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness? *Learning and Individual Differences*, 18, 430-445.

Hertzog, C., Price, J., Burpee, A., Frentzel, W., Feldstein, S., & Dunlosky, J. (2009).

Why do people show minimal knowledge updating with task experience:

Inferential deficit or experimental artifact? *The Quarterly Journal of Experimental Psychology*, 62, 155-173.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: support for a two-factor

theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 484-494.

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory.

Journal of Memory and Language, 48, 811-825.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization

approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349-370.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic

regulation of memory accuracy. *Psychological Review*, 103, 490-517.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The

production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning Memory and Cognition*, 36, 671-685.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the

generation effect: A three-factor theory. *Journal of Memory and Language*, 27, 521-536.

McDaniel, M. A., & Geraci, L. (2006). Encoding and retrieval processes in

- distinctiveness effects: Toward an integrative framework. In R. R. Hunt & J. B. Worthen (Eds.) *Distinctiveness and Memory* (pp. 65-88). Oxford: University Press.
- McNamara, D. S. & Healy, A. F. (1995). A procedural explanation of the generation effect: the use of an operand retrieval strategy for multiplication and addition problems. *Journal of Memory and Language*, 34, 399-416.
- Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 1263-1274.
- Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science: The effects of generation, errors, and feedback. *Psychonomic Bulletin and Review*, 14, 225-229.
- Morris, C., Bransford, J., & Franks, J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16, 519-533.
- 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.
- Postman, L., Burns, S., & Hasher, L. (1970). Studies of learning to learn: X. Nonspecific transfer effects in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 9, 707-715.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-209.

- Sahakyan, L. & Delaney, P. F. (2003). Can encoding differences explain the benefits of directed forgetting in the list-method paradigm? *Journal of Memory and Language*, 48, 195-201.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207-217.
- Slamecka, N. J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 592-604.
- Snedden, D. (1931). Practice Effect. *Journal of Educational Research*, 24, 376-382.
- Tiede, H. L., & Leboe, J. P. (2009). Metamemory judgments and the benefits of repeated study: Improving recall predictions through activation of appropriate knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35, 822-828.
- Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, 16, 465-47.

Appendix 1: Study Paragraphs

The target words appear in bold (read/generate)

Paragraph 1: Bloom's Taxonomy of Instructional Objectives

1. Blooms/ taxonomy of instructional objectives has been influential in education.
2. One set of classification is for the cognitive domain.
3. The lowest level of the cognitive learning objective is the **knowledge/kn-wl-dg-**level.
4. The next level is comprehension which includes **interpretation/int-rpr-tat-on**.
5. Application involves using principles of **abstraction/abstr-ct-on** to solve novel or real-life problems.
6. The **analysis/an-lys-s** objective involves having students see the underlying structure of an idea.
7. The **synthesis/s-nth-s-s** objective involves using skills to create completely new products
8. The most complex objective is evaluation, requiring value against some **criterion/cr-ter-on**.
9. Teachers can assure that their objectives cover many levels by writing a behavior content **matrix/m-tr-x**.
10. This shows how a particular concept will be taught and **assessed/ass-ss-d**.
11. Learning facts and **skills/sk-lls** is not the only important goal of instruction.
12. **Affective/Aff-ct-ve** objectives can be just as important.

Paragraph 2: Motivation and Goal Orientation

1. Some students are motivationally oriented toward learning goals.
2. However, some students are more oriented toward performance goals.
3. Learning goals are sometimes called **mastery/mast-ry** goals.
4. Students with learning goals see the purpose of learning as gaining **competence/c-mp-t-nc-** in skills being taught.

5. Students with performance goals primarily seek to gain positive **judgments/j-dgm-nts** of their competence.
6. Students with learning and performance goals do not differ in their **intelligence/int-ll-g-nc-**.
7. But they do differ in their **attitude/att-t-de**.
8. Learning-oriented students are more likely to use **metacognitive/metac-gnit-ve** strategies.
9. Learning-oriented students are also more likely to use self-**regulated/reg-lat-d** learning.
10. Performance-oriented students who perceive their **abilities/ab-l-ties** to be low are more likely to feel helpless.
11. Use of tasks that are **meaningful/me-n-ngful** will be more likely to lead to the adoption of learning goals.
12. Also, tasks that are **challenging/ch-ll-ng-ng** will be more likely to lead to a learning goal.

Appendix 2: Fill-in-the-blank Tests

Memory Test for: Bloom's Taxonomy of Instructional Objectives

Instructions: you will have 2 minutes to recall as many of the studied words as possible.

The lowest level of the cognitive learning objectives is the _____ level.

The next level is comprehension, which may include _____.

Application involves using principles or _____ to solve novel or real-life problems.

The _____ objective involves having students see the underlying structure of an idea.

The _____ objective involves using skills to create completely new products.

The most complex objective is evaluation, requiring value judgments against some_____.

Teachers can assure that their objectives cover many levels by writing a behaviour content _____.

This shows how a particular concept will be taught and _____.

Learning facts and _____ is not the only important goal of instruction.

_____ objectives can be just as important.

Memory Test for: Motivation and Goal Orientation

Instructions: you will have 2 minutes to recall as many of the studied words as possible.

Learning goals are sometimes called _____ goals.

Students with learning goals see the purpose of learning as gaining _____ in skills being taught.

Students with performance goals primarily seek to gain positive _____ of their competence.

Students with learning and performance goals do not differ in their _____.

But they do differ in their _____.

Learning-oriented students are more likely to use _____ strategies.

Learning-oriented students are also more likely to use self _____ learning.

Performance-oriented students who perceive their _____ to be low are more likely to feel helpless.

Use of tasks that are _____ will be more likely to lead to the adoption of learning goals.

Also, tasks that are _____ will be more likely to lead to a learning goal.

Appendix 3: Post Experiment Questionnaire

Instructions: Please think about each question carefully before making a response. If you cannot answer a question, please leave it blank.

1. What did you notice about how you did on the test on the **first** paragraph?

2. a. What strategy did you use to try to remember the **complete** items (e.g., kitten) on the **first** paragraph?

b. What strategy did you use to try to remember the **incomplete** items (e.g., k-tt-n) on the **first** paragraph?

3. Did you use a different strategy when studying the **second** paragraph? (Circle One)

YESorNO

a. What strategy did you use to try to remember the **complete** items on the **second** paragraph?

b. What strategy did you use to try to remember the **incomplete** items (e.g., k-tt-n) on the **second** paragraph?