BETTENCOURT, KATHLEEN M., M.A. An Investigation of the Test-Impaired New Learning Effect with Associative Recognition. (2016)
Directed by Dr. Thanujeni Pathman. 42 pp.

The test-impaired new learning (TINL) effect occurs when immediately after a test of old information, new information is presented to the learner. The effect, discovered by Finn and Roediger (2013), is that the new information is not remembered as well as if it were given to the learner after restudying the old information. This effect is important because it is counter to the well-established finding that testing leads to better recall than restudy. In the present research, we investigated the robustness of the TINL effect across two experiments that were identical except the type of test at final recall (associative recognition or cued recall). Participants studied word pairs (e.g., dog spoon), and during a second phase were tested on half the pairs and restudied the other half. Immediately after either a test or restudy trial, a new item was added to each pair (e.g. box), forming a triplet. In Experiment 1a (associative recognition test) participants were to identify pairs of old items as either same or rearranged. In Experiment 1b (cued recall test), participants were given the first item of the triplet and were to recall the second and third items. In Experiment 1a, we found improved memory for the old information after testing compared to restudy (testing effect); we found no TINL effect. In Experiment 1b, no testing effect was found, but there was evidence of TINL. This study adds to the small literature on the negative effects of testing and could have implications for both theoretical accounts of TINL and future applied work.

AN INVESTIGATION OF THE TEST-IMPAIRED NEW LEARNING EFFECT WITH ASSOCIATIVE RECOGNITION

by

Kathleen M. Bettencourt

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
2016

Approved by

_____
Committee Chair

APPROVAL PAGE

This thesis written by Kathleen M. Bettencourt has been approved by the following committee of the Faculty of The Graduate School at the University of North Carolina at Greensboro.

Committee Chair_____

Committee Members_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I thank my advisor, Dr. Thanujeni Pathman for her consistent guidance and immense patience throughout this process. I also thank my committee members, Drs. Peter F. Delaney and Michael J. Kane for their expertise and help on this project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Despite the intense aversion students at all levels feel toward formal examinations, testing is an effective method for improving memory. Testing can be adopted as a study strategy by quizzing yourself on the material for an upcoming exam. Years and years of research in cognitive psychology provide evidence that testing a person on studied information results in better memory for it (i.e., a higher exam score) on a later test (for reviews see Delaney, Verkoeijen, & Sprigel, 2010; Roediger & Karpicke, 2006a; Roediger & Butler, 2011). The *testing effect* is one of many well-supported findings included in a recent article that connects basic research findings related to study strategies with succeeding in college (Putnam, Sungkhasettee, & Roediger, 2016).

Testing someone on a random list of words can even increase memory performance for a second list of words studied *after* the test of the first list (for review, see Pastötter & Bäuml 2014). The forward influence of testing is seen with experiments where people learn multiple word lists. When participants take a test on a list before studying the next one, recall rates improve through a reduction of proactive interference—reduced performance/recall rates on test of List 2 (Szpunar, McDermott, & Roediger, 2008). Tulving and Watkins (1974) also observed a decrease in proactive interference as a consequence of testing with memory for word pairs. Participants learned

a list of A-B pairs (e.g., dog-spoon), and then took a test on them. After the test, they learned A-C pairs (e.g., dog-box). On a final test of A-C pairs, there was less interference than in the conditions without a test on A-B. The evidence here supports the idea that testing improves the ability to learn new associations (A-C) after being tested on old ones (A-B).

Potential Negative Effect of Testing

The studies discussed above are mere highlights of decades of research findings that paint testing as a beneficial method for learning. But this was called into question recently when Finn and Roediger (2013) discovered that recall of new pieces of information (e.g., words) was worse when presented immediately after test versus after restudy. Davis and Chan (2015) coined it *test-impaired new learning* (TINL), and replicated it in several experiments.

The basic paradigm included participants learning three-item sets of information—like face-name-profession, face-trait-profession, or word triplets (e.g., A-B-C- items). For example, in the experiments involving word triplets, during an initial study period, semantically related word pairs (e.g., shorts-pants, lettuce-salad) are presented for 5s each. Next, the word pairs are all either tested or restudied. The test trials had a cued recall format, where the first word of a pair (A-item: shorts) appeared on the screen. Participants were prompted to recall the second word (B-item: pants). After responding, they received feedback—the correct, intact A-B pair on the screen for 2s. In the restudy condition, the word pairs were all presented a second time for 5s. Immediately after the test or restudy of each pair, a new word (C-item: shirt) appeared underneath the A-B pair.

The instructions were to learn both the pair (A-B) and the new word (C), creating a list of A-B-C triplets. The last part of the experiment was a cued recall test. Like in the test condition, the A-item was presented on screen, but participants were asked to recall both the B ('old') and C ('new') items of the triplet in no particular order.

Across seven experiments, Finn and Roediger (2013) found that recall of B-items on the final test was higher in the test than restudy condition, a typical testing effect. But recall for C-items was significantly lower in the test condition than the restudy condition. This occurred with and without feedback on the intermediate test, and with semantically related and unrelated item triplets. The findings also replicated when the final test took place after a 24-hour delay. These results conflict with the research showing a positive forward effect of testing 'old' items on memory of 'new' items, namely the results of Tulving and Watkins (1974).

Theoretical Accounts of TINL

Finn and Roediger (2013) did not present a formal theoretical model for impaired new learning post-testing, but they briefly discussed a potential context-based explanation. They examined how performance on the initial test (successful vs. unsuccessful) influenced final recall accuracy, and found that unsuccessful test trials seemed to be responsible for TINL. Based on this, they speculated that both restudy and successful retrieval of the B-item on the intermediate test create a similar context—one that differs from unsuccessful retrieval on the intermediate test. The context of restudy and successful retrieval on the intermediate test then closely match the initial study phase context where participants originally saw the A-B pairs. The divergent contexts from the

intermediate phase can influence whether or not updating occurs—successful learning or incorporating of the new item into the set as evidenced by final test performance. As discussed by Finn and Roediger (2013), a more recognizable context is more likely to result in updating because it is recognized as being more familiar and thus better subject to slight changes. Slight changes in context here would be the incorporation of the new item into the context during the intermediate phase (test or restudy), which would result in better memory for the all item that make up the triplet. With successful intermediate retrieval trials and restudy, the match in context results in participants being better able to encode the new item. For unsuccessful retrieval trials, presumably the original study context is not remembered, which results in a lack of encoding of the new item (i.e., since the context is not recognized, it cannot be updated with additional information). From this view, it is the performance on the initial test that ultimately determines whether or not the new item will be successfully learned (i.e., the context will be updated).

Davis and Chan (2015) were the first to follow up Finn and Roediger ‘s (2013) study. They replicated the original TINL finding with face-name-profession triplets, and also extended the research with several additional experiments. They proposed the *borrowed-time hypothesis*, a metacognitive explanation for the occurrence of TINL. According to this account, the initial test plus new learning part of the experiment creates an encoding environment where participants focus more time on trying to learn the tested item (B-item) at the cost of learning the new item (C-item). This borrowing of time from one item to relearn another is said to occur because testing results in participants

considering the tested item as being either: important and thus should be remembered, likely to be test again in the future, or challenging to remember.

Several key findings provided evidence to support Davis and Chan's (2015) account. First, they found, similar to Finn and Roediger (2013) that in multiple experiments TINL seemed to be driven by unsuccessful initial test trials. Second, when the new learning is blocked separately from the test trials they found new learning was enhanced instead of impaired post-testing. Third, when the given study time during the initial test and new learning phase was increased, the magnitude of TINL was reduced. Lastly, when they gave participants an incentive for learning the new items TINL did not occur. Davis and Chan (2015) also emphasized that the advantage for learning of new items post-restudy was more due to how participants in the test condition were encoding information—engaging in metacognitive monitoring of whether or not they could correctly recall the B-item—than an advantage for restudy. On restudy trials, participants do not realize how difficult it would be to recall the B-item and thus probably are not worried about trying to relearn it. Thus, during restudy trials, participants do not borrow time from learning the new, C-item to relearn the B-item. Davis and Chan's (2015) account centers around metacognitive judgments about how well learned an item is, not just purely on their initial test performance.

Examining the Testing Effect and TINL with Associative Recognition

Given the overwhelming evidence in favor of testing for learning and memory, getting to the bottom of how and when testing has a positive versus negative effect on learning can help bolster a theoretical mechanism and inform more applied work on

classroom learning and teaching strategies. The present research seeks to investigate whether or not TINL occurs when the final test has an associative recognition format. In order to examine TINL in more applied settings, the use of recognition tests could be beneficial given that is a commonly used test format in academic settings (e.g., multiple choice tests).

All previous studies on TINL used cued recall as the final test (Finn & Roediger 2013; Davis & Chan, 2015). With cued recall, the A-item is given, and participants are to recall both the B- and C-items. Associative recognition tests, liked cued recall, are a method used to evaluate memory for pairs or associations. If the pairs from the experiment are AB, CD, EF, during the test participants see a mix of intact or old pairs (AB) and rearranged or new pairs (AD). The task is to label each pair as either 'old' or 'new'.

Previous work has shown that free and cued recall are similar to associative recognition via the recruitment and execution of the same retrieval processes. Specifically, word frequency effects show a distinction between single item recognition and associative recognition but no distinction between associative recognition and recall. Clark (1992) showed that with high frequency words, test performance is better for associative recognition and recall. But with low frequency words, performance is better for single item recognition. In regards to the process of retrieval, Nobel and Shiffrin (2001) regarded the similarity in reaction times (RTs) for cued recall and associative recognition as evidence that both test formats lead participants to engage in a sequential search of memory. Thus, cued recall and associative recognition are arguably equal

means of measuring memory for pairs or associations given the similarity of processes involved.

Although previous work on TINL used cued recall as the final test format, there has been work done on the testing effect with recognition tests (i.e., differentiating between old and new items). Most research on the testing effect includes free recall (i.e., participants are asked to write down all items from a list they studied earlier, no cues are given) as both the intermediate and final test formats (Chan & McDermott, 2007), but there is work investigating the effects of testing on final recognition tests. The literature on testing and recognition includes mixed findings.

Chan and McDermott (2007) investigated the effects of free recall testing on various types of final recognition tests in order to understand how testing influences the processes involved in recognition as outlined by the dual-process framework (Jacoby, 1991). The two processes involved are recollection—controlled, conscious and often involves the retrieval of contextual details of an event—and familiarity, which is characterized as more automatic and a sense of knowing something was seen previously without memory for specific details. With multiple experiments that examined source memory, exclusion, and remember/know judgments for single item recognition for words from two lists, Chan and McDermott found that participants in the testing condition (received immediate free recall test after the study of each list) had significantly higher recollection rates than those from the no-test condition. Participants in the no-test condition relied more on familiarity during final recognition. They concluded that testing

enhances the use of recollection processes on a later recognition test, which in turn produced better performance.

Two studies looked at the effects of retrieval on recognition with the retrieval practice paradigm, which examines memory for related word pairs from different categories. In the paradigm, participants first study word pairs during an initial study phase (e.g., banana-apple, shark-lion). Next, participants take a cued recall test that includes some of the word pairs from each category, but not all categories are tested here. Last, there is a final test on the word pairs. Verde (2004) used associative recognition as the final test, and found a testing effect—final recognition performance was better for pairs that were tested earlier in the study versus those that were not. With a single item recognition test—old and new words appear one at a time and are identified as old or new—Hicks and Starns (2004) also found higher scores for the previously tested items.

In a study investigating false memory, Roediger and McDermott (1995) found that an intermediate free recall test on a list of semantically related words produced better recognition memory of those later. Additionally with forced choice recognition—old and novel items are presented together, and participants pick out the old item—the benefit for items from an intermediate free recall test was also found (Hanawalt & Tarr, 1961). Based on these studies featuring a mix of intermediate recall and final recognition test formats, there does seem to be an overall benefit of testing. Chan and McDermott (2007) breakdown the above studies that produced mixed findings on the effects of testing on recognition within the dual-process framework. They suggest that the benefit of testing on recognition occurred when participants were unable to rely heavily on familiarity.

This includes single item recognition tests that include distractor items that are semantically related to target items (Hanawalt & Tarr, 1961; Hicks & Starns, 2004; Roediger & McDermott, 1995) and associative recognition tests (Verde, 2004).

Darley and Murdock (1971) looked at effects of intermediate free recall tests on lists of semantically unrelated nouns on final free recall and forced choice recognition tests. They found a testing effect on the final free recall test, but not on the recognition test. In a study examining serial position effects and influence of a recall test on later recognition, Jones and Roediger (1995) found no substantial benefit of intermediate testing on a later single item recognition test of low frequency words. With these studies, Chan and McDermott (2007) infer that neither recognition test required recollection-based processes and could be done by mostly relying on familiarity.

A study by Glover (1989) examined the effects of free recall, cued recall, and recognition intermediate testing on final test performance for the three test formats. In the experiments, participants read a short essay about a fake country during a first session. Two days later, participants returned and had some kind of test on the information, and four days after the first session they took a final test on the same material. The consistent pattern that emerged across experiments was that intermediate free recall tests (given two days after the first encoding session) had a significantly larger enhancement effect on final test performance on all final test types than both cued recall and recognition intermediate tests. Although the enhancement was greatest for free recall, cued recall and recognition intermediate tests also produced a significant benefit on the final test compared to a control group that received no intermediate testing. Taken together, it

appears that the majority of work on testing and recognition reveals a beneficial effect like with the copious amount of research on testing and recall (Roediger & Karpicke, 2006).

The Present Research

In order to investigate TINL with associative recognition as the final test, we adapted the Finn and Roediger (2013) paradigm. The stimuli consisted of sets of three words (e.g., dog spoon box, cabin student merchant, princess forest paper). In this paradigm, the participants studied A-B word pairs (e.g., dog spoon, cabin student, princess forest), then were tested on half of the word pairs and restudy the other half. During the test and restudy trials, C-items (e.g., box, merchant, paper) were interleaved with test and restudy trials, not blocked separately. A-B pairs were presented side by side and C-items appeared underneath the pairs. At final test, 'same' pairs (e.g., A-B: dog spoon, A-C: dog box) and 'rearranged' pairs (e.g., A-B: cabin forest, A-C: cabin paper) were presented. Participants identified each pair as either 'same' or 'rearranged'. To our knowledge, this is the first study that investigated TINL with associative recognition. We also included a second experiment that used the traditional paradigm's cued recall final test in order determine whether we could replicate both the testing and TINL effects with a new stimuli set of unrelated word triplets.

According to Finn and Roediger's (2013) context explanation for TINL, performance on the intermediate test will largely determine the degree to which TINL occurs. For pairs where participants correctly retrieve the B-item during the intermediate test, the C-item will be more readily learned. When the B-item is not retrieved, a revision

to the context is less likely to be made leading to TINL for those triplets. Given the larger number of triplets and semantically unrelated nature, it is less likely that participants will perform well on the majority of the intermediate test items. With a majority of the intermediate test trials resulting in unsuccessful retrieval, it is probable that TINL will be apparent.

From the perspective of Davis and Chan's (2015) *borrowed-time hypothesis*, the change in final test format from cued recall to associative recognition will not affect the occurrence of TINL. Based on their account's foundation being the encoding environment of mixed test and new learning trials, they would predict that the same impairment in learning of C-items will appear as a result of focusing on relearning the B-items in the test condition.

Based on the similarities between cued recall and associative recognition (Clark 1992; Nobel & Shiffrin, 2001) and the replicated TINL finding with cued recall, we predicted similar findings for the two experiments. For the first experiment (Experiment 1a) that included an associative recognition final test, we predicted a testing effect for the A-B pairs and a TINL effect for the A-C pairs when looking at hits (correctly identified 'same' pairs). More specifically, we expected that participants would be significantly better at recognizing A-C pairs learned following restudy than test. According to Chan and McDermott's findings on testing's influence on performance for recollection-based tests, a testing effect should emerge for the A-B pairs. For the second experiment (Experiment 1b) that included a cued recall final test, we predicted a replication of both

the testing and TINL effects—higher performance on B-items from test condition and higher performance on C-items from restudy.

CHAPTER II

EXPERIMENTS 1A AND 1B

Method

Both Experiments 1a and 1b had the same initial and intermediate encoding phases, but the final test format was different. In Experiment 1a, we used an associative recognition final test, and in Experiment 1b, we used a cued recall final test.

Participants

A total of 64 (32 in Experiment 1a, 32 in Experiment 1b) young adults from the University of North Carolina Greensboro participated in the study, and received course credit for doing so. Participants were randomly assigned to be in either Experiment 1a (M_{age} = 19.50, *SD* = 1.48) or 1b (M_{age} = 19.39, *SD* = 1.50). Demographics for the two studies were closely matched with 8 males and 24 females in each experiment.

Materials

The stimuli consisted of 120 English nouns words from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982) that vary in length from four to eight letters, and have concreteness ratings greater than or equal to four on a seven point scale ranging from abstract (1) to concrete (7). These words were randomly selected from the 199 nouns that had concreteness ratings of four or more. The 120 words were put into a random order, and sorted into 8 study and 8 test lists once for all participants. We created 40 word triplets (A-, B-, and C-item sets), with only the exception that all words in a triplet were semantically unrelated.

If any triplets included related words, one or more of the words were switched out with semantically unrelated words from the list.

Half of each study list consisted of 20 triplets that would be in the test condition, and 20 triplets that would be in the restudy condition during the second phase of encoding. Thus, with the creation of 8 study lists, all 40 triplets had an equally likely chance of being in either the test or restudy condition during phase two of encoding. To create each test list, each study list of 40 triplets was sorted into a matrix to determine how the triplets would be tested during a final associative recognition test. The test lists consisted of 40 A-B and 40 A-C pairs, where the A-item for all pairs was the same that participants saw during encoding (‘old’), but the B- and C-items were either the same (forming a ‘same’ pair) or rearranged (forming a ‘rearranged’ pair). The test lists each consisted of 20 ‘same’ A-B pairs, 20 ‘rearranged’ A-B pairs, 20 ‘same’ A-C pairs, and 20 ‘rearranged’ A-C pairs. All A-B and A-C pairs had an equally likely chance of being ‘same’ or ‘rearranged’, with the 8 different lists. Each study list was composed of 120 words (40 A-B-C triplets), and each test list was composed of 160 words (40 A-B pairs and 40 A-C pairs).

Procedure

Participants were randomly assigned to 1 of 8 possible counterbalancing orders. They were told that they would be learning word pairs (A-B), and later on in the experiment they would learn a new word (C-item) to go with each pair. Participants were not told anything about how they would be tested, but that they would be tested on which words were shown together. The first two phases of the experiment consisted of encoding

of stimuli, and the third/last phase was an associative recognition memory test on A-B and A-C pairs.

During Phase 1 of encoding, participants in all conditions saw 40 word pairs (the first two words A-B of each A-B-C triplet) on a computer screen in black Arial font on a white background. Each pair was presented side-by-side in the center of the screen for 5s. See *Figure 1* for experimental procedure diagram. The presentation order of the pairs was randomized across all counterbalances. After all 40 pairs were presented, participants worked on a 1-minute math distractor task, in which they solved addition problems with paper and pencil.

During Phase 2 of encoding, participants in all counterbalances were tested on 20 of the word pairs (A-B) from the first phase. The first word of each pair (A-item) appeared on the screen with a question mark next to it, and participants were prompted to recall the second word of the pair (B-item) by typing it into a text box on the screen. After typing their response, participants were instructed to press enter. If participants could not recall the word, than they were instructed to just press enter. There was no time limit on this test. After pressing enter (whether a response was entered or not), the correct pair (A-B) from the first phase appeared on the screen for 2s. After 2s a new/third word (C-item of the A-B-C triplet) was presented underneath the A-B pair. The triplet was displayed on the screen for 5s.

The other 20 word pairs (A-B) from the first phase were restudied. For the restudy pairs, each pair was presented on the screen for 5s, followed by presentation of a new/third word (C-item of the A-B-C triplet) underneath the A-B pair. The triplet was

displayed for 5s. The order of the test and restudy trials was mixed and randomized in all counterbalances. After the second phase, all participants worked on a word search for 10 minutes with paper and pencil.

Final Test Phase: Experiment 1a. Phase 3 of the experiment for all counterbalances consisted of an associative recognition test of two blocks, one for A-B pairs and one for A-C pairs. The order of the recognition blocks was counterbalanced. Participants received general instructions stating that they would be seeing pairs of words on the screen, and they were to identify each as either 'same' by pressing 1 on the keyboard, or 'rearranged' by pressing 2 on the keyboard. To ensure participants did not forget the instructions, the key labels were present at the bottom of the screen for each recognition trial. For the A-B block, the instructions stated that 'same' pairs were A-B words that were presented together during phases one and two, 'rearranged' pairs are pairs where both A- and B-items were presented during phases 1 and 2, but were not paired together. The instructions for the A-C block were identical. Both recognition blocks (A-B and A-C) were self-paced.

Final Test Phase: Experiment 1b. Phase 3 of Experiment 1b, across counterbalances, consisted of a cued recall test. Participants were given the A-item of each triplet, and were to recall both the B- and C-item that were paired with the cue previously. They provided their answers by typing them into a response box on the screen. Instructions were to recall the two items in any order.

Results

The intermediate test accuracy for Experiments 1a (M = .24, SD = .16) and 1b (M = .21, SD = .19) were not significantly different from each other $t(62) = .64$, $p = .52$. The range of intermediate test performance for each experiment (1a: 0-65%; 1b: 0-70%) was also similar. The lack of difference here reinforces that the experiments were the same except for the final test format.

Experiment 1a

Associative Recognition Measures. *Table 1* includes means and standard deviations of hits and false alarms and d-prime values (*d'*) for the A-B and A-C Blocks by condition (Test vs. Restudy). Hits are correctly identifying same pairs as same, and false alarms are incorrectly identifying rearranged pairs as same. The measure of d-prime is from signal detection theory (Egan, 1958; Stanislaw & Todorov, 1999). The theory applies to tasks where participants are to label something most basically as old or new (e.g., a 'same' pair would be considered old and an 'rearranged' pair would be considered new). The d-prime values provide a measure of sensitivity when a person is classifying a pair as either 'same' or 'rearranged'. Higher d-prime values indicate that a person is better at distinguishing between same and rearranged pairs. Lower d-prime values reveal that a person was not quite as good at differentiating between the two kinds of pairs. If d-prime values are either 0 or negative, it is normally indicative of either sampling error or response confusion (Stanislaw & Todorov, 1999). In addition, we calculated criterion-c for the A-B and A-C blocks to determine if participants were biased toward responding

'same' or 'rearranged'. The formula for criterion-c is: [-0.5*(z-score for Hits) + (z-score for False Alarms)].

Recognition Memory for Pairs. A 2 (Item type: A-B vs. A-C pairs) x 2 (Condition: Test vs. Restudy) repeated measures ANOVA with proportion of Hits as the dependent measure was conducted. There was a significant main effect of item type, $F(1, 31) = 95.07$, $MSE = 2.82$, $p < .01$, $\eta_p^2 = .75$, showing that overall A-B same pairs ($M = .87$, $SD = .11$) were better recognized than A-C same pairs ($M = .57$, $SD = .19$). This is not surprising given that participants saw each A-B pair twice during the experiment, and only were exposed to each A-C pair once. The main effect of condition was not significant, $F(1, 31) = .17$, $MSE < .01$, $p = .68$, $\eta_p^2 = .01$, which reveals that across A-B and A-C pairs, there was no overall effect of test versus restudy on final associative recognition accuracy. But the interaction between item type and condition was significant, $F(1, 31) = 6.26$, $MSE = .14$, $p = .02$, $\eta_p^2 = .17$.

A follow-up paired t-test revealed a significant difference in recognition of A-B pairs by condition, $t(31) = 2.52$, $p = .02$, $d = .45$. A-B 'same' pairs from the test condition were better recognized than those from the restudy condition. In other words, a testing effect was found for the A-B pairs. For A-C 'same' pairs, the follow-up paired t-test did not show significant differences, $t(31) = -1.45$, $p = .16$, $d = .26$, between conditions. Participants' recognition of A-C 'same' pairs did not significantly differ for the test versus restudy conditions. Although here there is not a significant TINL effect for associative recognition, additional evidence is needed in order to form a conclusion about TINL and this test format. It is possible that additional studies with larger sample sizes

could produce the effect. Also, it is interesting to note that testing did not produce an enhancement or facilitation effect on new learning on the associative recognition test.

False Alarms. A 2 (Item type: A-B vs. A-C) x 2 (Condition: Test vs. Restudy) repeated measures ANOVA was done to examine effects on false alarms. There was a significant main effect of item type, $F(1,31) = 36.42$, $MSE = 1.14$ $p < .01$, $\eta_p^2 = .54$, which showed that overall there were more false alarms made on A-C pairs ($M = .43$, $SD = .17$) than on A-B pairs ($M = .24$, $SD = .17$). Participants were more likely to mistake an A-C 'rearranged' pair as being 'same' than for an A-B pair. There was no main effect of condition, $F(1,31) = .04$, $MSE < .01$ $p = .84$, $\eta_p^2 < .01$, meaning that across test and restudy participants did not differentiate in their false alarm rates. The interaction between item type and condition was not significant $F(1,31) < .01$, $MSE < .01$ $p = .96$, $\eta_p^2 < .01$.

d-prime Analyses. A 2 (Item type: A-B vs. A-C pairs) x 2 (Condition: Test vs. Restudy) repeated measures ANOVA with d-prime values as the dependent variable was conducted. There was a main effect of item type, $F(1,31) = 125.37$, $MSE = 75.29$, $p < .01$, $\eta_p^2 = .80$, showing that d-prime values were much higher for A-B pairs ($M = 2.01$, $SD = .90$) than for the A-C pairs ($M = .48$, $SD = .82$). This difference indicates that participants were better at correctly recognizing A-B pairs as either 'same' or 'rearranged' than A-C pairs. The main effect of condition was not significant, $F(1, 31) = 1.50$, $MSE = .31$, $p = .23$, $\eta_p^2 = .05$, revealing overall, across A-B and A-C pairs there were not differences in d-prime for test versus restudy. The interaction between item type and condition approached significance, $F(1, 31) = 3.82$, $MSE = .92$, $p = .06$, $\eta_p^2 = .11$.

Follow-up t-tests revealed a significant difference in d-prime values for A-B pairs in the test condition (M = 2.15, SD = .89) versus restudy condition (M = 1.88, SD = 1.00), $t(31) = 2.57$, $p = .02$, $d = .45$. Like with the analysis using proportion of hits as the dependent measure, participants better recognized A-B pairs from the test condition than those from restudy. Concerning A-C pairs, there was not a significant difference between those from the test (M = .44, SD = .97) versus restudy condition (M = .52, SD = .83), $t(31) = -.55$, $p = .60$, $d = .10$. Thus, the pattern of results from the analysis with hit proportions of A-B and A-C pairs paralleled the findings using d-prime values.

Criterion-c. To determine if participants' were biased toward responding 'same' or 'rearranged' for A-B and A-C pairs, criterion-c for each block was calculated. Typically more negative values indicate a bias toward responding 'same', and more positive values indicate a bias toward responding 'rearranged'. For the A-B block, participants had a slight bias toward responding 'same' (M = -.20, SD = .30), and for the A-C block, no bias was apparent (M = .01, SD = .35) since the value was so close to zero.

Reaction Time Analysis. As an exploratory analysis, a 2 (Item Type: A-B vs. A-C pairs) x 2 (Condition: Test vs. Restudy) repeated measures ANOVA was conducted with mean reaction times of hits as the dependent measure. The goal was to see if the same pattern was present in the hit reaction times for A-B and A-C pairs. There was a significant main effect of item type, $F(1, 30) = 37.85$, MSE = 25260760.26, $p < .01$, $\eta_p^2 = .56$, which revealed that across conditions hit reaction times for A-B pairs (M = 1694.46, SD = 527.18) were significantly faster (lower) than for A-C pairs (M = 2597.16, SD = 1012.31). The main effect of condition was also significant, $F(1, 30) = 6.05$, MSE =

3196574.34, $p = .02$, $\eta_p^2 = .17$, and showed that across item type participants' hit responses were faster on pairs from the test ($M = 1985.25$, $SD = 617.53$) than the restudy condition ($M = 2306.37$, $SD = 920.05$). The interaction between item type and condition was not significant, $F(1, 30) = .87$, $MSE = 2258979.22$, $p = .36$, $\eta_p^2 = .03$.

Intermediate Test Performance and its Effect on Final Test. *Table 2* includes means and standard deviations of final test performance for cases when the B-item was correctly versus incorrectly recalled during the intermediate test. Intermediate test performance was examined in relation to final recognition performance for both the A-B and A-C pairs. The was done to determine whether or not correctly retrieving a B-item during the intermediate test gave participants an advantage in recognizing the pair as either 'same' or 'rearranged'.

For A-B pairs, a paired t-test revealed no significant difference in overall correct recognition of A-B pairs $t(31) = 1.75$, $p = .09$, $d = .31$, when performance on the intermediate test was correct ($M = .91$, $SD = .23$) versus incorrect ($M = .83$, $SD = .15$). The same pattern emerged for the A-C pairs with no significant difference in overall correct recognition of A-C pairs $t(31) = .1.91)$, $p = .07$, $d = .34$, when performance on the intermediate test was correct ($M = .66$, $SD = .31$) versus incorrect ($M = .55$, $SD = .18$). For A-B and A-C pairs, performance on the intermediate test does not appear to significantly influence final test performance. We can cautiously speculate that with similar future studies there might be an effect on intermediate on final test performance. But like with TINL and associative recognition, more evidence is needed in order to move toward a firmer conclusion.

Summary. Using an associative recognition final test, we found that for 'old learning' (A-B pairs), participants performed significantly better in the test condition than the restudy condition. This difference is seen in both the hit proportion and d-prime scores. However, we did not find a TINL effect for the A-C pairs using either hits or d-prime scores. Participants made more false alarms on A-C pairs than on A-B pairs. D-prime values for the A-C pairs were also much lower overall than for the A-B pairs. D-primes lower than zero often indicate misunderstanding of instructions or a task being done incorrectly (e.g., pressing 'same' for a rearranged pair and 'rearranged' for a same pair regularly) (Stanislaw & Todorov, 1999). In this case, it could also be that the difficulty of the task contributed to these extremely low d-prime values.

Performance on rearranged pairs showed only that overall, participants were better at identifying A-B versus A-C rearranged pairs. We did not find evidence that participants' accuracy on the intermediate test affected final recognition performance. But the pattern of results did indicate an advantage on the final test for both A-B and A-C pairs where the B-item of the triplet was correctly retrieved on the intermediate test.

Experiment 1b

Final Cued Recall Performance. A 2 (Item type: B-item vs. C-item) x 2 (Condition: Test vs. Restudy) repeated measures ANOVA was conducted to examine final test performance. There was a significant main effect of item type, $F(1,31) = 91.74$, $MSE = 2.96$, $p < .01$, $\eta_p^2 = .75$, which revealed that overall participants recalled more B-items ($M = .41$, $SD = .27$) than C-items ($M = .11$, $SD = .16$). The advantage for B-items is most likely due to the repeated presentation of each B-item during the second phase of

encoding. The main effect of condition was not significant, $F(1,31) = 1.90$, $MSE = .02$, $p = .18$, $\eta_p^2 = .06$, meaning that across B- and C-items, there was no effect of test versus restudy. But there was a significant interaction between item type and condition, $F(1,31) = 17.17$, $MSE = 11$, $p < .01$, $\eta_p^2 = .36$.

The pattern of results can be seen in *Figure 2*. Follow-up paired t-tests revealed no significant difference $t(31) = 1.26$, $p = .22$, $d = .22$, between recall of B-items for the test versus restudy condition. Final recall performance of B-items was not different for those that were tested versus restudied. For C-items though, recall was significantly different between conditions, $t(31) = -3.99$, $p < .01$, $d = 70$. Participants' recall of C-items was better for the restudy than for the test condition. Thus, although no testing effect is observed with the B-items, test-impaired new learning is apparent for the C-items.

Intermediate Test Performance and its Effect on Final Test. *Table 3* shows means and standard deviations of final test performance for cases when the B-item was correctly versus incorrectly recalled during the intermediate test. Intermediate test performance was examined in relation to recall of both B- and C-items. For B-items, final recall was significantly better when intermediate test performance was correct ($M = .96$, $SD = .12$) versus incorrect ($M = .35$, $SD = .21$), $t(28) = 14.79$, $p < .01$, $d = 2.75$. The same pattern was observed for C-items, $t(28) = 2.41$, $p = .02$, $d = .45$, where final recall of C-items was significantly better if a participant correctly recalled the B-item for the triplet during the intermediate test ($M = .15$, $SD = .26$) versus if they incorrectly or did not recall it ($M = .04$, $SD = .09$).

Summary. Experiment 1b replicated the TINL effect with C-items found by both Finn and Roediger (2013) and Davis and Chan (2015). A testing effect for the B-items did not appear, but it is not uncommon for testing effects to be absent on tests given after short delays (Roediger & Karpicke, 2006a). Although we did not find an effect of testing, there does seem to be a large advantage for final recall of both B- and C-items if the B-item of the triplet is correctly retrieved during part 2 of encoding. Overall recall rates were lower than in previous studies examining TINL, but that is most likely due to the increased number of stimuli (e.g., previous work used 20 triplets, and the current study used 40). The words used here were also semantically unrelated, whereas past work used related word triplets (Finn & Roediger, 2013).

There was a large difference in recall rates of both B- and C-items conditional on intermediate test performance. Correctly recalling the B-item of a pair on the intermediate test significantly increases the chances of recalling not only that item on the final test but also the C-item.

CHAPTER III

DISCUSSION

Test-impaired new learning (TINL) is a relatively new phenomenon discovered by Finn and Roediger (2013). The basic effect is that introducing "new" information for someone to learn immediately after they are tested on old information results in poor later memory for the "new" item. The poor later memory piece is illustrated by comparing recall of new items that were presented immediately after a test on an old item versus after an old item is studied a second time. Previous studies investigating TINL used cued recall as the final test format (Roediger 2013; Davis & Chan, 2015).

When we examined TINL with associative recognition (Experiment 1a), there was not a significant TINL effect for A-C pairs. Additional work with this paradigm and test format is needed for a solid conclusion to be reached. The difference between the mean hit rate for test and restudy did follow the typical TINL pattern—better performance for A-C pairs from the restudy condition—so it is possible that with a larger sample the effect could potentially be present for associative recogntion. There was a testing effect for A-B pairs—participants performed better on correctly recognizing A-B pairs from the test condition than from the restudy condition. The effects were found both with proportion of hits and d-prime values as dependent measures. We included a second version of the experiment (Experiment 1b) that maintained the cued recall final test format used in the original Finn and Roediger (2013) study, with the extension that both a

different and larger set of stimuli (unrelated word triplets) was used. The opposite pattern from Experiment 1a was found for Experiment 1b—TINL was replicated for the C-items, but there was no testing effect for B-items. Recall rates were lower than in previous studies, but this makes sense given the increase in to-be-remembered items.

Intermediate test performance was examined for both experiments to determine if successful compared to unsuccessful retrieval influenced final test performance. In Experiment 1a, the difference in final recognition test performance based on the intermediate test did not reach significance for either A-B or A-C pairs. But the pattern did show a slight advantage for pairs where the B-item was successfully retrieved on the intermediate test. In Experiment 1b, final recall of B- and C-items was significantly better when the B-item was successfully retrieved on the intermediate test. The latter finding is similar to previous work on TINL (Davis & Chan, 2015; Finn & Roediger, 2013).

For Experiment 1a, an exploratory analysis looking at the effects of item type and condition on RTs for hits found that across condition, participants made hit responses faster for A-B pairs. Also, across item type participants made hit responses quicker for those from the test condition. These findings could indicate that participants spent more time on test trials during the intermediate phase since they were self-paced. The extra time spent on the trials might influence participants to feel more confident and respond faster to pairs from those specific trials. An interesting follow-up study could use a recognition test that evaluates participants' memory for the source (test vs. restudy trial) as well as whether the pairs are same or rearranged. If participants have more accurate source judgments for test trials, it could mean they spend more time on them. This could

also be interpreted as test trials being more salient to participants given the reasons discussed by Davis and Chan (2015) and thus easier to remember. Assessing source memory could also provide a way in which to examine recollection and familiarity processes present on the final test.

Implications for Theoretical Accounts

The findings from Experiment 1a do not rule out either the context account or the borrowed-time hypothesis. Finn and Roediger's (2013) context account of TINL is based on intermediate test performance and its influence on whether or not the new item is learned. With both successful retrieval and restudy participants see the pair a second time, which mirrors the original study context. The closeness of these contexts then makes it more flexible for updating—incorporating of a new item to the set. Based on the increased amount of triplets and the nature of the items in each not being semantically related, it was probably that intermediate test performance would be poor. With more unsuccessful retrieval trials, it then was likely that TINL would occur with associative recognition as it has with cued recall. In Experiment 1a, testing did not significantly impair new learning, and the intermediate test performance did not significantly influence final recognition. At first glance it seems then that the context account might not be the best explanation, but the pattern of results still closely resembled TINL both overall and in looking at the intermediate test performance's effect on final recognition. The performance on the intermediate test was low and did not differ from Experiment 1b where TINL was found. It is possible that context does play a role, but not in the exact way that Finn and Roediger (2013) discussed it.

A transfer appropriate processing (TAP) account would not predict TINL for an associative recognition test, since that format more closely matches both the initial study phase and intermediate phase for A-B and A-C pairs, respectively. The TAP account is based on findings that memory on a final test is better when the retrieval situation or context is the same as that of initial studying (Morris, Bransford, & Franks, 1977). With TINL studies, a cued recall test does not match up as well with the learning environment of the new items as an associative recognition test. On an associative recognition test, participants are shown both an A-item and C-item, just as they were during the intermediate phase for restudy and test conditions. From a TAP perspective, it is possible that the match and mismatch between test formats and the intermediate phase in the experiments is why a significant TINL effect was found for cued recall but not for associative recognition. This interpretation is cautious though, given this is the first study examining TINL with associative recognition.

Davis and Chan's (2015) *borrowed-time hypothesis* also predicted a TINL effect for associative recognition, given that it focuses on the intermediate phase with mixed test and new learning trials as the catalyst for the impairment post-testing. A non-significant difference in A-C recognition between test and restudy conditions might mean that there is more to the effect than just the encoding environment. Presumably, participants' metacognitive monitoring does affect which items they choose to focus on learning, but without self-reports on strategy use or learning prioritization, the extent of the effect is not completely clear. Future work could incorporate a measure looking at such things, and potentially strengthen the case for Davis and Chan's (2015) account.

Potential Differences for Associative Recognition

Given that cued recall and associative recognition are said to rely on the same processes (Nobel & Shiffrin, 2001), this finding was unexpected. In order to conclude that TINL is or is not present on associative recognition tests, additional work with larger samples is needed. More experiments could replicate the current study—no TINL effect for associative recognition—or find an enhancement of new learning post-testing. Replications of Experiment 1a's findings would call into question the similarity in retrieval processes used during cued recall and associative recognition tests.

Experiment 1a's findings add to the mixed literature on effects of testing on later recognition tests. Chan and McDermott (2007) determined that testing produces more recollection-based processes at retrieval, which can lead to better recognition memory performance compared to a no testing condition. Given that associative recognition is arguably a mostly recollection-laden task (Yonelinas, 2002), it would seem that testing would produce an enhancement of memory for both the A-B and A-C pairs. But the current work did not compare testing to a no testing condition but rather a restudy condition. Thus, differences in how test and restudy uniquely contribute to recollection processes need to be considered here.

Verkoeijen, Tabbers, and Verhage (2011) followed up Chan and McDermott's (2007) work and compared the influence of both testing and restudying on recollection in an exclusion recognition memory test. When initial test performance was relatively low (~50%), they found that testing and restudying produce about the same amount of recollection at final test. Additionally, they found that when feedback was presented

during test and when the test was made easier, testing resulted in higher rates of recollection than restudy. Across all experiments, restudy produced higher levels of familiarity than testing.

The current study's procedure includes a mix of components from Verkoeijen, et al. (2011), with lower initial test performance (<30%) and the incorporation of feedback during test trials. It could be that in Experiment 1a, the testing and restudy conditions resulted in relatively similar levels of recollection during final test for the A-C pairs, hence no significant difference between the two. The significant testing effect for A-B pairs potentially indicates that more recollection was produced as a result of testing the A-B pairs than restudying them. Additional work investigating TINL with associative recognition could better examine the unique contributions of test and restudy on recollection at final test and might illuminate more about potential differences in retrieval processes produced by test versus restudy.

Future Directions

The overall better performance on both recognition of A-B pairs and recall of B-items aligns well with the *list strength effect*—items that are strengthened are better remembered. The repeated presentation of A-B pairs in the Experiment 1a and B-items in Experiment 1b can be referred to as a strengthening of those items. Because participants saw A-B pairs or B-items a second time, their memory for them was better than the A-C pairs or C-items that were only seen once for a few seconds. Similarly, the overall low recall rates observed in Experiment 1b could be a consequence of the *list length effect*, which shows that increasing the number of items in a list decreases memory performance

on that list (Ratcliff, Clark, & Shiffrin, 1990). Hence, the performance on the final test here was much lower than in previous studies using 20 triplets where all items were semantically related (e.g., lettuce salad tomato).

Final recognition and recall of A-C pairs and C-items was relatively poor overall. For recognition, there were a relatively large number of participants with d-prime scores low enough to indicate no ability to differentiate between same and rearranged pairs. With cued recall, about one-third of the participants did not remember a single C-item. These low performance rates are indicative of the task's high level of difficulty. Like with A-B pairs and B-items there's also the same consideration of pure list length. The number of A-C pairs or C-items to-be-remembered combined with the short presentation time may start to explain the almost floor level of performance.

Performance on the final test did span a relatively large range for both pair types (A-B Hit proportion: .65-1.00; A-C Hit proportion: .15-.85). These ranges could indicate that there is an individual difference effect in the sample. Based on the ranges in performance, a direction this research could go in at the basic level is to determine if there is an individual difference effect present, which could help explain the distribution that ranges from high to floor performance levels. By including some measure of individual differences, it would be possible to get at who in the sample is really causing the effects. In a within-subjects replication of Finn and Roediger's (2013) Experiment 6, Bettencourt and Delaney (under revision) found a negative relation between the testing effect and TINL. It appeared in their study that the people showing a testing effect were not showing as much of the TINL on C-items. This inverse relation between the two

effects would support the idea that not the same people are contributing to the two different effects.

In order to examine the unique contributions of test and restudy on later retrieval processes, additional work could include exclusion recognition tests and ones that require source judgments about whether a pair came from restudy or test could help explain the occurrence of TINL. Perhaps if within this specific paradigm restudy and test produce similar levels of recollection and familiarity, this is an instance where testing does not have an enhancement effect on future learning. Coinciding with Davis and Chan's (2015) view, this is largely influenced by the mixing of test trials with new learning, so maybe when those are blocked separately testing could increase recollection for A-C pairs on a final test more so than restudy. The inclusion of measures that evaluate confidence at final test, strategy use during encoding, and individual differences could also shed more light on why and when TINL is present. The use of recognition tests to examine TINL is also useful for work that might examine the underlying neural processes of recollection and familiarity during final test (for review see Rugg & Curran, 2007).

Research on TINL could also go in a more applied, hands-on direction in order to bridge the gap between this basic cognitive psychology research and classroom learning. Additional experiments using stimuli that are semantically related is a step in that direction, given that students are usually not taught to remember a series of concepts or ideas that are not associated with each other. This effect has potential implications for when teachers ask students questions in class. If the student responds incorrectly or does not know the answer, the teacher (or another student) could provide the correct answer

plus some additional related information. Perhaps later on an exam, the student's memory for that additional information might be weaker. In order to determine this, the paradigm used in this study and others to study TINL could be adapted to work in a classroom. One possible method would be to have students have either a brief quiz or review (restudy) of material from the previous class. After each quiz question or restudy trial, a new but related concept from the current lesson could be presented. At the end of a semester, students' memory for those 'new' concepts could be assessed in order to see if TINL is observed in this more practical setting.

Conclusion

In the first study examining test-impaired new learning with a recognition test, specifically associative recognition, the effect did not reach conventional levels of significance, but the pattern in favor of TINL was apparent. There was a testing effect for the A-B pairs on the recognition test, but for both A-B and A-C pairs there did not appear to be a significant advantage for correct recall of B-items during the intermediate test on final recognition performance. Again, the pattern of successful retrieval positively influencing later memory was there for both A-B and A-C pairs, but the differences were not significant. The cued recall test version of the experiment did produce the TINL with a new set of materials, but no testing effect was found. Here, unlike with the associative recognition test, there was a relatively large and significant advantage for final recall of both B- and C-items when the B-item of the triplet was correctly recalled during the intermediate test.

Future studies should consider incorporating individual difference measures to understand more about whether or not the effects are caused by the same people. Additional work with recognition as the final test can include measurements of recollection and familiarity in order to understand how test and restudy contribute to those within this paradigm. Adapting the encoding environment by separating restudy and test from new learning trials with a final recognition test might also provide more evidence for Davis and Chan's (2015) *borrowed-time hypothesis* as a theoretical mechanism of TINL.

Research on TINL could also be extended to a more applied domain like classroom learning and teaching strategies. By doing this, the potential for TINL in practical settings could be understood and strategies developed to prevent it. Since most tests in education settings are more similar to a recognition format, continuing this research with different test formats is imperative. An ultimate goal of this research is to bridge findings from basic research with those from a practical domain. If similar effects are found both in the lab and the classroom, then additional work in both settings could help to illuminate ways to enhance students learning opportunities.

REFERENCES

Bettencourt, K.M., & Delaney, P.F. (under revision). A close look at test-impaired new learning: Disentangling the costs and benefits.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 431-437.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition, 20*(3), 231-243.

Darley, C.F., & Murdock, B.B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91*(1), 66-73.

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: how does testing impair new learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Delaney, P.F., Verkoeijen, P. J. L., & Sprigel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In: B. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (63-147). San Diego, CA: Elsevier Academic Press.

Egan, J.P. Recognition memory and the operating characteristic. Bloomington, Indiana: Univer., Hearing and Communication Laboratory, Technical Note AFCR-TN-58-51, 1958.

Finn, B., & Roediger, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1665-1681.

Friendly, M., Franklin, P.E., Hoffman, D., & Rubin, D.C. (1982). The Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation 14*(4), 375-399.

Glover, J.A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399.

Hanawalt, N. G., & Tarr, A. G. (1961). The effect of recall upon recognition. *Journal of Experimental Psychology, 62*(4), 361-367.

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review, 11*(1), 125-130.

Jacoby, L. L. (1991). A process dissociative framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541.

Jones, T. C., & Roediger, H. L. (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology, 7*(1), 65-80.

Morris, C.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 384-413.

Pastötter, B., & Bäuml, K. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology, 5*.

Putnam, A.L., Sungkhasettee, V.W., & Roediger, H.L. (2016). Optimizing learning in college: Tips from cognitive psychology. *Perspectives on Psychological Science, 11*(5), 652-660.

Ratcliff, R., Clark, S.E., & Shiffrin, R.M. (1990). List strength effect: Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179-195.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181- 210.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Rugg, M.D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*(6), 251-257.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Test during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392-1399.

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on recall of another. *Journal of Verbal Learning and Verbal Behavior, 13*, 297-309.

Verde, M. F. (2004). The retrieval practice effect in associative recognition. *Memory & Cognition, 32*(8), 1265-1272.

Verkoeijen, P. P. J. L., Tabbers, H. K., & Verharge, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology, 58*(6), 490-498.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

APPENDIX A

TABLES AND FIGURES

Table 1

Means and Standard Deviations of Hits, False Alarms, and Mean d-prime for AB and AC Pairs

	Test			Restudy		
	Hits	False Alarms	*d'*	Hits	False Alarms	*d'*
AB Pairs	.91 (.11)	.23 (.21)	2.15	.83 (.16)	.24 (.20)	1.88
AC Pairs	.54 (.23)	.43 (.20)	.52	.60 (.20)	.43 (.19)	.44

Table 2

Means and Standard Deviations of Final Test Accuracy for A-B and A-C Pairs and B- and C-items from Successful versus Unsuccessful Intermediate Recall for Experiments 1a and 1b

	Successful Recall	Unsuccessful Recall
1a: A-B Pairs	.91 (.23)	.83 (.15)
1a: A-C Pairs	.66 (.31)	.55 (.18)
1b: B-items	.96 (.12)	.35 (.21)
1b: C-items	.15 (.26)	.04 (.09)

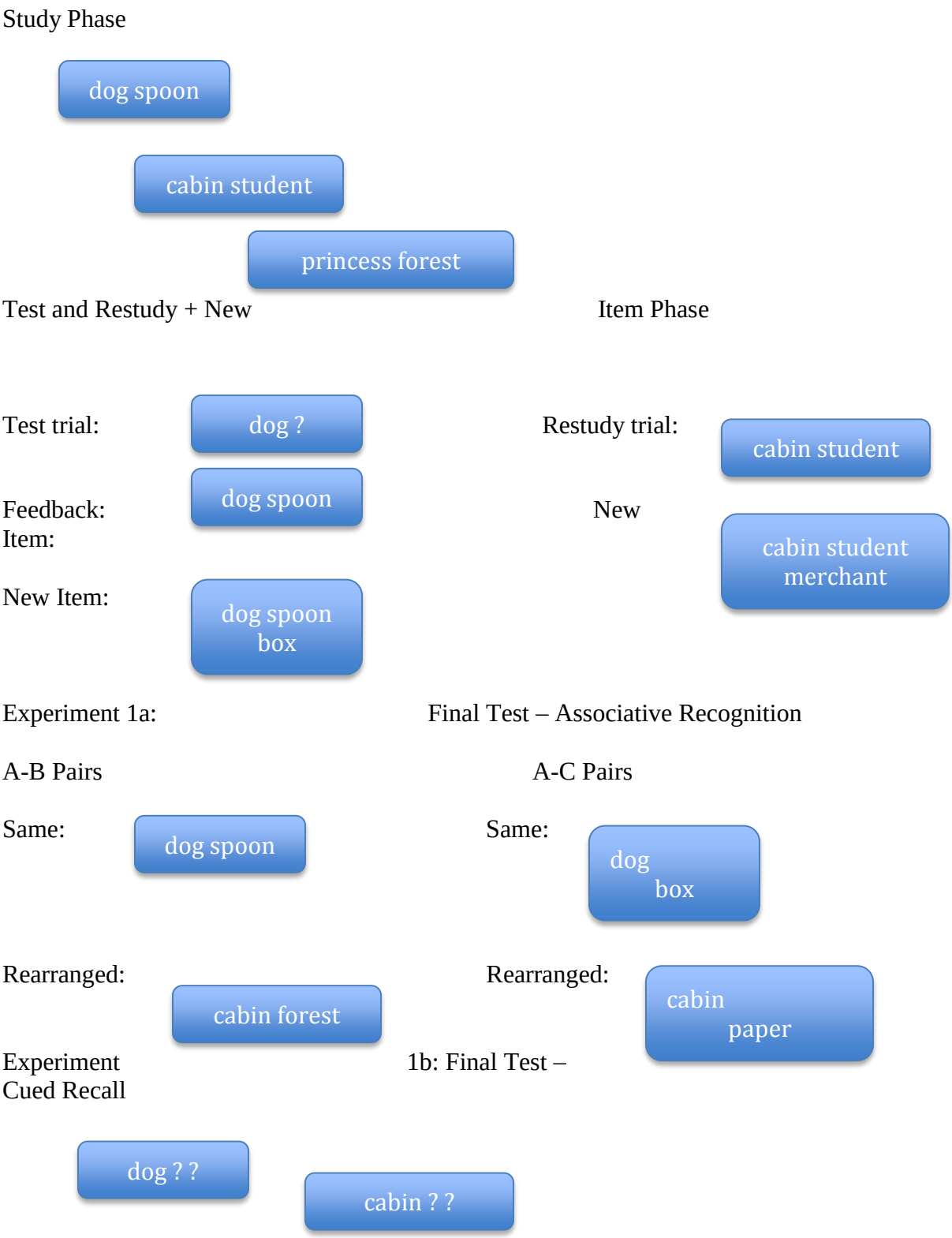

Figure 1. Paradigm of Experiments 1a and 1b.

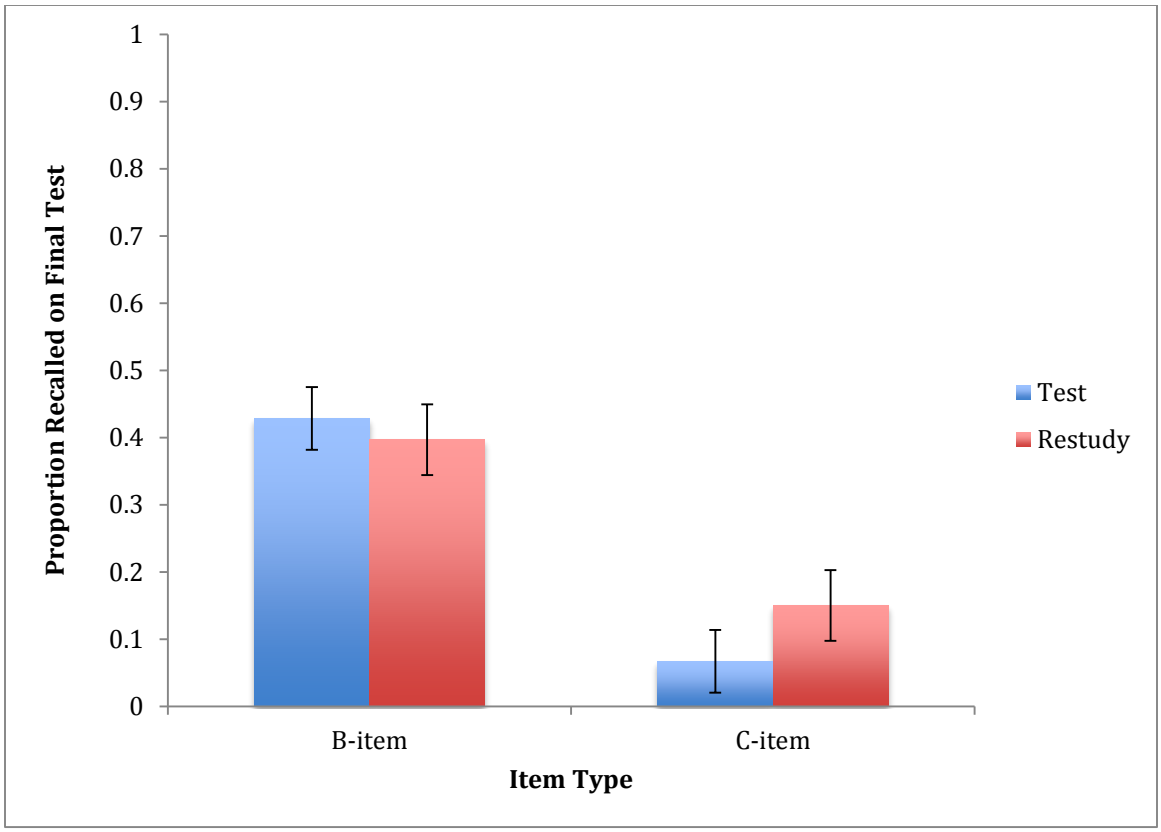

Figure 2. Mean Proportions of B- and C-items Recalled for Experiment 1b.