Overt and Covert Retrieval Effects on Memory for Repeated and Novel Word-Pairs

by

Monique Carvalho

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements
for the degree of

Masters of Science

in

Psychology with a collaborative specialization in Neuroscience

Guelph, Ontario, Canada

ABSTRACT

OVERT AND COVERT RETRIEVAL EFFECTS ON MEMORY FOR REPEATED AND NOVEL WORD-PAIRS

Monique Carvalho Advisor:

University of Guelph, 2019 Harvey Marmurek

This thesis examined the impact of retrieval type, either overt or covert, compared to restudying on the recall of word-pairs in a forward testing paradigm. Participants were presented with a list of word-pairs (A-B) before completing one of three tasks: (1) overt retrieval; (2) covert retrieval; or (3) restudying. Participants then studied a novel list (C-D) or an interference list (A-D) before completing cued-recall of that list. Two experiments were conducted: (1) between-subjects design; (2) mixed design where task varied between-subjects, but item type varied within-subjects. Experiment 1 found that whereas covert retrieval produced forward testing effects for both lists, overt retrieval only trended towards producing an effect for the A-D list. In both experiments A-D items produced more intrusions than C-D items. Both retrieval methods reduced intrusions. The results are discussed in relation to integration theory of the forward testing effects.

Keywords: memory and learning, testing effect, forward testing effect, overt retrieval, covert retrieval

Acknowledgements

I would like to acknowledge and thank Dr. Harvey Marmurek for all of his valuable constructive feedback during the planning and development of this work, and all the time he has given to helping me. I would like to thank and acknowledge Dr. Christopher Fiacconi for his time and feedback. I would also like to thank the Marmurek lab and its past students for all of their help in data collection. I would like to thank the University of Guelph for the resources provided to me during my time of study. Finally, I would like to thank my family for all of their support and encouragement throughout my study.

Table of Contents

List of Tables

List of Figures

List of Appendices

Overt and Covert Retrieval Effects on Memory for Repeated and Novel Word-Pairs

Learning and memory are important for everyday lives and especially in a school setting. Almost everyone has had the experience of having thought that they knew the information that they would be tested on but has been surprised on the test day just how little they can remember. This limitation on memory performance may be informed by an understanding of the testing effect. The testing effect is a phenomenon in which having previously been tested on material improves later recall of that material relative to repeated studying (Mulligan, Susser & Smith, 2016; Rickard & Pan, 2017; Rowland, 2014; Sundqvist, Mäntylä, & Jönsson, 2017). The testing effect has been demonstrated with various materials: video lectures; single words; paired-associate words; and, written passages (Yang, Potts, & Shanks, 2018). The testing effect has also been tested across a wide range of participants, including an online sample, college students and individuals with brain damage (Chan, Meissner & Davis, 2018). A variation of the testing effect is the forward testing effect (Pastötter & Bäuml, 2014). The forward testing effect is defined as better recall of material after having been tested on different material relative to repeated studying. In this thesis I will refer to the effect of testing on learning a single set of materials as the testing effect in contrast to the effect of testing one set of materials on the learning of a new set of materials as the forward testing effect. Alternate terminology distinguishes between test enhanced potentiation of relearning and potentiation of new learning (Chan et al., 2018).

The Testing Effect

Studies of the testing effect (Buchin & Mulligan, 2017; Cho, Neely, Crocco &Vitrano, 2017; Endres & Renkl, 2015; Racsmány, Szőllősi, & Bencze, 2017; Rickard & Pan, 2017; Rowland, 2014) typically involve three phases: (1) a study phase during which participants encode the stimuli, (2) a review phase, and (3) a test phase. During the review phase, participants

are presented with either a memory test for the previously studied items or a restudying phase. Those tasks may be varied between subjects or within subjects. In place of a review participants may be presented with a distractor task. The final test phase involves a memory task for the material presented during the study phase.

The Forward Testing Effect

The forward testing effect is defined as improved memory for new material following testing of prior material (Cho et al., 2017). In an early study of the forward testing effect (Tulving & Watkins, 1974) participants studied a list of unrelated word pairs followed by a picture. Participants were either tested (cued-recall) on that list or drew the picture. Participants then studied a second list of paired-associates followed by a picture (see Figure 1 for an overview of their procedure). The word-pairs in the second list contained the stimuli from the first list but they were paired with new words. That is, the two lists conformed to an A-B, A-D negative transfer paradigm. On an immediate cued recall test of the second list mean correct recall was 45% if the first list had been tested and was 15% if the first list was not tested. Tulving and Watkins (1974) proposed that "explicit [overt] retrieval of stored information seems to insulate the A-B list from the A-C [A-D] list in a way that removes the former as an interfering component in the learning of the latter" (p. 192).

Cho et al. (2017) compared the magnitudes of the testing and forward testing effects (see Figure 2 for an overview of their procedure). Participants studied a list of 16 Swahili-English word-pairs followed by a filler math task that lasted for 28 seconds. During the review phase participants were either tested (cued-recall test) on the word-pairs or restudied the word-pairs. There were two successive review cycles for each group. Participants then studied a new list of 32 word-pairs consisting of the 16 old pairs and 16 new pairs where the new pairs did not contain

any of the old items. The second list was repeated for two cycles. Cued-recall for the second list was better if the first list had been tested rather than restudied. Although overall recall was better for old than for new items, the benefit of testing was equivalent for old items (18% testing effect) and new items (21% forward testing effect). That it, testing had a general benefit rather than an item-specific benefit.

The Cho et al. (2017) study differs from the Tulving and Watkins (1974) study in several critical procedures. Whereas Tulving and Watkins (1974) used an A-B, A-D relationship between lists, Cho et al. (2017) used a new list that mixed repeated (A-B, A-B) items and completely novel items (A-B, C-D). Thus, it is not clear whether the benefits of testing would be item-specific in the negative transfer condition. Another difference between the studies is that Tulving and Watkins (1974) did not include a restudy baseline condition. Rather, they used the picture drawing condition as a baseline which may lead to an overestimate of the benefits of testing. One aim of the present thesis was to compare the effects of testing on old and new items where the new items varied in terms of their susceptibility to item-specific interference. For that purpose, both completely new (A-B, C-D) and negative transfer (A-B, A-D) items were studied along with repeated items (A-B, A-B) items and the review phase manipulation included a restudy condition.

There are various factors that may impact the testing effect and forward testing effect (Chan et al., 2018): (1) between-subjects design studies yield larger forward testing effects than within-subjects design studies; (2) filler tasks such as solving math problems and no retrieval tasks produce larger effects than restudying the material; (3) related items between the first and second list (related old and new items) produce larger effects than unrelated material; (4) testing improves recall to a larger degree when the testing formats are changed from one set to the next

(the opposite pattern is found within intrusions); (5) testing effects are larger as the amount of material previously studied increases; (6) feedback during initial testing impairs new learning; and, (7) unlike in the testing effect where long retention intervals increase the effect, in the forward testing effect long retention intervals diminish the effect.

Theories of the Testing Effect and Forward Testing Effect

Theories of the testing effect and forward testing effect can be generally grouped into four larger categories, as described in Chan et al. (2018). These four accounts or theories are: (1) resource theories; (2) metacognitive theories; (3) integration theories; and (4) context theories. Integration and context theories apply only to the forward testing effect, since they require both old and new material. One theory alone may not account for effects.

Resource theories assume that prior testing potentiates new learning by increasing the cognitive resources that are available to the individual at the time of encoding (Chan et al., 2018). That is, being tested on List 1 items frees up space in memory so that new information can be learned. If participants restudy the material, that material is still sitting in memory waiting to be used. The relative gain in resources can be accomplished either by increasing the amount of attentional resources available during the subsequent encoding, or by reducing proactive interference. Proactive interference occurs in studies where participants are presented with an interference second list (same cue as the first list, now paired with a novel target). Testing in these conditions helps to reduce or eliminate the number of intrusions (target items from a prior list being mistakenly recalled as target items from the new list) compared to restudying (Chan et al., 2018). In this way testing helps to insulate encoding-based proactive interference.

Metacognitive theories attribute the testing effect and forward testing effect to learners optimizing their encoding strategies (Chan et al., 2018). This improvement in the way that

participants learn may be due to a change in the strategy they use to learn the information. In comparison to resource theories that assume that without testing there is a decrease in the amount of resources available for encoding, metacognitive theories assume that testing teaches the individual to improve their encoding strategies. This may be because without prior testing participants are often overconfident in their ability to recall information, or because being tested may increase an individual's perceived likelihood of being tested again.

An example of this metacognitive theory can be shown in the idea that the testing effect should be larger as the level of difficulty of retrieval increases (e.g., increasing the time between initial learning of the material and testing, or increasing the amount of information to be recalled). This relates to the idea that the testing effect is driven by retrieval efforts. The more difficult the task, the more likely the participant requires stronger retrieval strength to correctly recall the information. Bjork and Bjork's (1992) theory of disuse states that memories are described in terms of their strength in storage and retrieval and that the less an item is retrieved the less likely it will be able to be recalled at a later point. Storage strength is related to how well a memory is formed; retrieval strength relates to how easily the memory is able to be retrieved. Rickard and Pan (2017) suggested that the testing effect may occur due to increased retrieval strength rather than improved storage strength. The main differences between the encoding or retrieval strengths and efforts, are that the 'effort' is placed upon the individual attempting to either encode or retrieve the information more effectively, while the 'strength' refers to the individual's ability to recall or encode the information, and how those strengths can be weakened without use. Although longer times between tasks improve recall in the testing effect, this is not the case for the forward testing effect. In order to make tasks in the forward testing effect other

techniques such as increasing the amount of information to be learned would need to be implemented.

Integration theories assume that the benefit of testing is due to enhanced integration of the new material with the old material (Chan et al., 2018). Testing increases the individual's ability to remember the old information when they are learning the new information, thereby binding the new and old information. This theory accounts for why the forward testing effect occurs within an interference (A-B, A-D) design. The original 'A-' cue becomes associated in memory with both the original '-B' target and the new '-D' target. This may be accomplished by triggering an updating system which allows the old information and new information to be integrated together into an existing memory, such that when the 'A-' cue is presened individuals are able to recall both the '-B' and the '-D' targets. This benefit of retrieval is linked to the participant's believed relevance of the learnt material. If participants believe the material to be relevant to their learning the learned material triggers an extended retrieval which encourages integration of the old and new information. Comparatively, if they deem the information irrelevant, the previously learnt material hinders the learning of the new material. The forward testing effect should be larger for A-D than C-D items if participants believe that the previously learned material is relevant to the new material being studied.

An example of integration theory can be seen in Wahlheim (2015) study. He tested whether interference was reduced by testing (see Figure 3 for his procedure). Participants initially viewed a list of word-pairs and were asked either to restudy items or recall items. Half of the pairs were restudied, and the other half were tested via cued-recall in a within-subjects design. The word-pairs were semantically related (target words were not related to one another). For restudy items participants read the pairings out loud. For tested items participants were

instructed to say the missing word out loud before the end of four seconds. Responses were written down by the experimenter.

Participants were then presented with a second list which included repeated A-B items, interference A-D items and control C-D items in a mixed list (Wahlheim, 2015). Participants were instructed to read the pairs out loud as they appeared on the screen. During the testing phase of the second list items, participants were asked to type out the missing word from each pair. Following each pair participants used a sliding scale to indicate whether the pairing had changed from the first list to the second list once recall of the item was completed. Wahlheim (2015) found that A-D items were recalled better if the items with those A stimuli were tested rather than restudied. There was no difference in recall between the A-D items and the C-D items for items tested following the first list. However, there was lower recall of the A-D items compared to the C-D items when those items had been restudied following the first list. Moreover, intrusion rates in the restudied A-D items were higher restudied than tested items. These findings demonstrate that proactive interference is mitigated by the testing of the first list.

Change recollection rates were higher for A-D items than C-D items (there should not be any change detected in the C-D items), and those rates were higher for items that had been tested compared to restudying. Change recollection rates for the interference condition were greater when there was correct recall than incorrect recall of the first list items. This may have led to a facilitation in the recall of the A-D items. Participants may have been better at detecting change for the A-D items than the C-D items because they were asked to recall part of the initial A-B list. The common stimulus would serve as a cue as to which word to expect on the second list. Overall, the results of the Wahlheim (2015) study indicate testing items on list one eliminates proactive interference leading to better recall of the A-D pairs, consistent with Tulving and

Watkins (1974). In Wahlheim's (2015) second experiment participants were provided the correct word-pairs as feedback during list one testing. Proactive interference in the A-D condition resulted in lower levels of recall than for the C-D items which was inconsistent with their first experiment results; however, change recollection rates still remained higher for the A-D items than the C-D items. This inconsistency in the results was attributed to the feedback given during the testing of the initial items.

Context theories assume that the benefits of testing derive from an individual's ability to distinguish old information from novel information (Chan et al., 2018). This is in contrast integration accounts. Testing creates a new context that helps to separate the old information into a context that is different from the encoding of the new information. The attempted retrieval of the information creates an internal context change that simply restudying the information does not. This is different from the resource theory in that creating the different contexts for the memories helps to reduce the size of the search-set that participants undergo to retrieve the desired information.

Another mechanism proposed to underlie the testing effect and forward testing effect is transfer appropriate processing (TAP). According to TAP, the testing effect and the forward testing effect derive from the similarity of processing engaged during the first and subsequent tests (Rowland, 2014; Veltre, Cho, & Neely, 2015). The retrieval processes engaged during the initial test are similar to those on the final test. Restudying alone does not engage the relevant retrieval process. TAP leads to the prediction that the testing effect should increase in proportion to the degree of process similarity between tasks (Rowland, 2014). TAP may drive the forward testing effect in that simply being exposed to the testing paradigms may increase recall during

list two by alerting participants to the format that the later task will use even though the items themselves may have changed.

Endres and Renkl (2015) found that the testing effect was not driven by TAP. Participants were presented with three passages, a distractor word-game task, and then a review phase. For each passage participants were tested on the content using one of three methods: restudied the passage; free recalled the passage; or completed a short-answer task. Participants completed each review task in a within-subjects design. After a one-week delay, participants completed free recall for all three passages, followed by a distractor word-game before completing six final short-answer tasks (three questions of which were initially tested, three of which were initially not tested) for one of the three passages. If TAP was driving the effect, then there would be an interaction between the initial task type and subsequent task type results.

Endres and Renkl (2015) found that there was no interaction between initial and subsequent task types. Recall was better when participants had either performed the short-answer task or free recall task than when they restudied the passage. For previously non-tested items, short-answer testing significantly outperformed the restudy condition. There was no difference between performing the free recall or the short answer tasks. These results do not support TAP theory. It is possible that this failure was due to the conditions used in the study. While recall of word-pairs and recall of the content of the passage task are both assessing short term episodic memory, the difference between having to recall general compared to specific information may be why they did not find evidence for TAP. A critique of passage recall is that the main idea is likely to be recalled and there is a reasonable flow of information from one point to the next. This allows for connections to be made more easily and remembered. In comparison, when the recall task requires remembering specific information, such as word-pairs, there may not be any

flow or connections between recalling the list of items. Similarly, recalling information in the short answer questions may cue the participants to other parts of the initial passage, and this is not something that would occur with word-pairs. In contrast, trying to recall a list of unrelated information is more challenging than recalling information that is seemingly interrelated. In relation to TAP, the processes for free recall and short-answer may be similar in that the information is interrelated, which may explain why they found that the type of test was not a significant factor.

Carpenter, Pashler and Vul (2005) also reported evidence against TAP. In their first experiment participants studied 40 weakly associated word-pairs in a within-subjects design. Participants then restudied 20 of the word-pairs and engaged in a cued-recall test for the remaining 20 word-pairs. In the cued-recall condition participants were presented with the 'A' term and had to recall the 'B' term. The word-pair was shown again following attempted retrieval. The next day participants completed one of four final recall tasks: (1) cued-recall where they were presented with the 'A' term and had to recall the 'B' term; (2) cued-recall where they were presented with the 'B' term and had to recall the 'A' term; (3) free recall of the 'A' terms; or, (4) free recall of the 'B' terms. Tested items produced greater recall than the restudied items regardless of the type of final test. Cued-recall produced larger testing effects than free recall. There was an interaction effect where testing produced a greater benefit in the recall of 'B' items compared to 'A' items in the cued-recall condition. Carpenter et al. (2005) stated that finding a testing effect for both types of cued-recall was evidence against TAP.

Comparing Overt and Covert Retrieval

Tulving and Watkins (1974) attributed the testing effect to an explicit retrieval process. The vast majority of testing effect studies employ overt retrieval defined as information that is

both retrieved and articulated either in written or verbal form (Sundqvist et al., 2017). However, recent studies have focused on whether covert retrieval leads to a testing effect. Covert retrieval of information can be defined as "an answer that is retrieved and produced internally by thinking it, but with no overt articulation of that information" (Sundqvist et al, 2017, p. 2). One challenge for studies of covert retrieval is to operationalize the covert retrieval process.

Previous literature has found that covert retrieval produces levels of recall similar to overt retrieval. Sundqvist et al. (2017) compared overt and covert retrieval in a testing effect paradigm. Participants underwent the typical three phases of a testing effect. In the first phase participants studied word-pairs three times, with the cycles separated by a 30 second mathematics task. During the review phase participants completed four separate tasks (covert retrieval, write, type, and restudy). Word-pairs were evenly divided across each condition in a mixed list within-subjects design. In the restudy condition participants were presented with the word-pairs again. In the retrieval conditions a cue word was presented to the participant who attempted to recall the word that was missing from the pair. Participants were asked to press the ENTER key if they believed that they knew the missing word. If the ENTER key was not hit, the next cue word was shown. If the participant did hit ENTER, they were prompted to either write down or type the word (overt retrieval) or to wait for the next word to appear (covert retrieval). The final test phase occurred after a delay of five minutes or seven days.

Sundqvist et al. (2017) found that participants pressed ENTER at similar levels for items pertaining to covert and overt (type and write) conditions. There were significant effects of retention interval, response format and the interaction between those factors. There was no testing effect at the short retention interval (five minutes). At the long retention interval (seven days), there was a testing effect. Cued recall performance was lowest in the restudy condition

and equivalent across the covert and write conditions. Across a series of experiments and a meta-analysis of previous studies, Sundqvist et al. (2017) found that overt retrieval led to a larger testing effect than covert retrieval in some specific conditions (e.g., the manner of the overt response). They concluded, however, that the testing effect is primarily the result of retrieval processes such that articulation adds little to what is produced by retrieval itself.

Overview of The Current Studies

The first goal of the present experiments was to study the effects of testing on old (A-B, A-B) and new items (A-B, A-D, or A-B, C-D). A second goal of the current studies was to compare overt and covert retrieval effects in the forward testing paradigm developed by Tulving and Watkins (1974). The comparison of overt and covert retrieval may have implications for the effectiveness of study habits by students who exercise covert retrieval rather than overt retrieval while reviewing to-be-tested material.

A 3 X 3 (Review [cued-recall, JOL, restudy]) x List 2 [A-B, A-D, or C-D]) design was implemented. Participants were presented with an initial list of word-pairs (A-B), following which they were presented with either an overt test, covert test, or restudy. Participants then viewed a second list of word-pairs that varied in relation to the first list pairs: (1) the cue words were the same as in list one but were paired with a new target word (A-D); (2) completely new word-pairs (C-D); or, (3) repeated word-pairs (A-B). An overt retrieval test was completed on the second list. In the first experiment (see Figure 4 for thesis procedure), a between-subjects design was used for the tasks completed following the first list (cued-recall, JOL, restudy) and for the list two type (A-B; A-D; C-D). In the second experiment, a between-subjects design was used for the tasks following presentation of first list (cued-recall, JOL, restudy) and a within-

subjects design was used for item type (A-B, A-B/A-D/C-D). The dependent variable was the number of correct target words recalled for the second list.

The first experiment in the current thesis utilized a completely between-subjects design in accordance with the seminal study of Tulving and Watkins (1974). The first experiment differed from the Wahlheim (2015) studies in that a between-subjects design was utilized instead of a within-subjects-design. Participants were subjected to the standard testing effect items (A-B, A-B), as well as two forward testing effect conditions: interference (A-B, A-D), and control (A-B, C-D) conditions. The review (restudy, overt recall, covert recall) tasks were also be varied between-subjects. In the second experiment, list conditions were varied within-subjects, and the review tasks were manipulated between-subjects as in Tulving and Watkins (1974). The reason that task was manipulated between-subjects was because Chan et al. (2018) found that between-subject effects produced larger testing effects and forward testing effects than within-subjects designs.

This thesis also aimed to compare overt and covert retrieval on the testing effect and forward testing effect. Tulving and Watkins (1974) stated that explicit retrieval was the source of the reduced proactive interference, but they did not examine a covert retrieval condition. This study is important because it compares overt and covert retrieval methods on the forward testing effect with completely novel stimuli (A-B, C-D) and with interference stimuli (A-B, A-D), while also examining their impacts on the testing effect. Additionally, this study aimed to determine whether TAP contributes to the forward testing effect (List 1: A-B, List 2: A-D, C-D), as well as the testing effect (List 1: A-B, List 2: A-B). It also aimed to determine if the effects were item specific or general (due to a generalized improvement in either encoding and retrieval, that can be seen in novel list, Cho et al., 2017).

The following hypotheses were tested (see Figure 5):

Hypothesis 1: The repeated A-B list would produce the greatest recall, followed by the C-D list and then the A-D list. It was hypothesized that the C-D list will produce greater recall levels than the A-D list due to proactive interference in the A-D list.

Hypothesis 2: The largest forward testing effect will occur for the A-D list because testing will reduce the number of intrusions that occur. That is, the difference between cued-recall retrieval and restudy will be the greatest for A-D items. This is expected under the integration theory since the A-D list shares the same cue as the first list. There will be a generalized benefit of testing for the C-D list. No testing effect is expected since participants will be completing the tasks without delays (Chan et al., 2018). Although Cho et al. (2017) found a testing effect with a brief delay between lists, they used multiple reviews for the first list and multiple presentations of the second list.

Hypothesis 3: If the testing effect is due to TAP, overt retrieval methods will produce a stronger testing effect and forward testing effect than covert testing methods because the retrieval mechanisms overlap across tests to a greater extent when both tests involve overt retrieval. Sundqvist et al. (2017) found some benefits of overt retrieval on the testing effect but concluded that retrieval in any form would produce a testing effect. However, they did not compare overt and covert retrieval for forward testing effects.

Experiment 1

Method

Material. The stimuli consisted of 144 word-pairs derived from the *English Lexicon Project* (Balota et al., 2007). The words were common concrete nouns ranging between three to

five characters in length that had a Thorndike-Lorge frequency of 25 per million or greater as in Tulving and Watkins (1974).

Paired associates were used because they were found to produce larger testing effect size estimates than using single words (Rowland, 2014), but more importantly to keep the experiment consistent to the one used in Tulving and Watkins (1974). The words were paired randomly and screened to ensure that there were no obvious semantic connections between the pairs. All word-pairs were subjected to a Latent Semantic Analysis where words that scored between -.15 and .15 were considered unrelated (Landauer & Dumais, 1997). Words that were pluralized or proper nouns were excluded. The word-pairs were subdivided into six lists containing 24 word-pairs per list: A-B, A-D, and C-D. Lists were created to form control (C-D), interference (A-D), and repeated (A-B) conditions (see Appendix A). In order to control whether the results were due to the specific word lists created, two versions of these lists were created.

The stimuli were projected onto a large screen. Both the cued-recall task and the JOL task were given to participants on paper, and the responses were written on the provided paper (see Appendix C). Response sheets were separated by an opaque dark page so that participants could not see through the first page onto the next. The response sheets were held together in a clipboard.

Participants. Tulving and Watkins (1974) reported that the effect of testing yielded $t(14)= 4.97$, for the difference between their immediate test condition and picture drawing condition, which is equivalent to a Cohen's *d* effect size of 2.66. G-power indicated an *N* of 26 per condition for a large effect size (.80) and power of .80. I tested 241 participants (there were 26-29 participants in each group) who were students at the University of Guelph, enrolled in a first-year psychology class. Participants were recruited through SONA, an online system that

allows its users to sign up for a study of the user's choice. Participants were granted a participation credit in their coursework.

Experimental Manipulations. Participants were assigned to experimental conditions according to the condition scheduled for their session. All participants first viewed a 24 A-B word-pair list. Participants were tested in one of the following nine groups: (1) overt retrieval of list one with final test A-D recall; (2) overt retrieval of list one with final test C-D recall; (3) overt retrieval of list one with final test A-B recall; (4) covert retrieval of list one with final test A-D recall; (5) covert retrieval of list one with final test C-D recall; (6) covert retrieval of list one with final test A-B recall; (7) restudy of list one with final test A-D recall; (8) restudy of list one with final test C-D recall; and, (9) restudy of list one with final test A-B recall (see Table 1). Each of the second lists contained 24 word-pairs.

Procedure. Participants were first presented with a consent form (see Appendix B for the study titled Studying to Remember) to read. Participants were encouraged to ask any questions they had pertaining to the consent form prior to signing it.

Participants were tested in groups of up to six people per group. Stimulus presentation was controlled by *PsychoPy* (Peirce, 2009). The researcher read aloud the instructions which were also presented on the screen, to the participants and answered any questions before moving on to the next step. Participants were welcomed to the study and told that they were going to be viewing 24 word-pairs followed by a picture. They were told that following the initial study period they would be asked to recall the word-pairs, judge their ability to recall the word-pairs, restudy the word-pairs, or draw from memory the picture that appeared at the end of the list (no participants were asked to draw). The picture was included to eliminate recency effects by having participants focus attention on the image. The software randomly presented each word-

pair for four seconds. At the end of the list a simple drawing was presented for 15 seconds. The simple drawing consisted of two stick figures interacting with the objects. Participants were informed about the duration of each word-pair and image.

As depicted in Figure 4, participants engaged in a practice trial in which they viewed five randomly created word-pairs followed by a picture (practice items were not used in any of the critical word-pair lists). The duration of the word-pairs and image presentations were the same as in the actual trials. Following the presentation of the list and image, participants were shown a sample of what they would complete if they were in the word-recall condition, JOL condition, restudy condition, or the picture drawing task. They were then given a chance to ask any questions they had before being presented with the first trial. Before the trial began participants were reminded that they would be viewing 24 word-pairs followed by an image and that following this presentation they would be asked to recall the word-pairs, judge their ability to recall the word-pairs, restudy the list, or draw the image.

Participants then viewed the first set of 24 word-pairs (A-B list) followed by an image. Following the end of the initial study period, participants were asked to flip over their clip board and follow the instructions on their sheet. They were told that they had two and a half minutes to complete the task.

Participants in the overt retrieval condition were presented with the first word (A- word) on the screen and were instructed to write down the word that was paired with it (-B word) on their response sheet (see Appendix C for an example of the response sheets they received) for all 24 word-pairs. No specific instructions were given for words they could not recall. Participants in the covert retrieval condition were presented with the first word (A- word) on the screen and were instructed to write down on their response sheet on a scale of 0 to 100 how strongly they

felt they could recall the missing word. Participants in the restudy condition were presented with all the word-pairs for a second time in a new random order.

In the review phase, each word-pair or cue word was presented on the screen for 6 seconds. Participants were instructed to restudy the word-pairs as they appeared on the screen if they were in the restudy condition, write down the missing word if in the cued-recall condition, or write down a number indicating their perceived likelihood of recalling the missing word if they were in the JOL condition. Once they completed the first task, participants were asked to take the first page and slide it face down underneath their clipboard, so that all they could see is the opaque dark page covering the final task. This was done to ensure that participants were focusing on the next task. Additionally, by having participants place their first task underneath their clipboard it ensured that on the next task they were not able to refer back to the previous task.

When all participants were ready, they were told that they were going to view a second list of 24-pairs of words followed by a picture. The second list was the repeated A-B list, A-D list, or C-D list. Participants were not told what the second list would be. They were reminded that each word-pair would be presented for four seconds and that the image would be presented for 15 seconds. Participants were instructed that they would either be completing a cued-recall task or drawing the image from memory. Following the presentation of the second list, participants were asked to complete a cued-recall task. Participants engaged in another practice trial following the same format as the first list. Participants again had a chance to ask questions before beginning the second task. Before beginning the next task, participants were reminded that they would view 24 word-pairs followed by a picture, and that the task following the picture was to recall the word-pairs or draw the image from memory. Participants were then presented

the second list in a random order. Once the list was finished, participants were instructed to remove the opaque paper and follow the instructions on the following sheet. All participants completed a cued-word recall where the cue word was presented on the screen in a random order and the participants made their responses on the response sheet provided for them. Each cue word was presented on the screen for six seconds.

Upon completion of the task, participants were given a debriefing form to keep which explained the full purpose of the study (see Appendix B). They were encouraged to contact the faculty researcher if they had any questions. After completion of the experiment, participants were granted their SONA credits.

Analysis. The second list recall accuracy was subjected to a 3 X 3 (Review [cued-recall, JOL, restudy] x List 2 [A-B, A-D, or C-D]) analysis of variance (ANOVA). A 3 X 2 (Review [cued-recall, JOL, restudy]) x List 2 [A-D or C-D]) ANOVA was conducted on the number of intrusions that occurred in the second list. A correlational analysis was run between average JOLs for the first task and final recall, as well as between number of items recalled correctly for List 1 and 2. For all analyses, alpha was set at 0.05. Bonferroni corrections were applied to any post-hoc comparisons. Cohen's *d* (effect size) is reported for all *p*'s less than .10.

Results

A one-way ANOVA of recall of List 1 words showed a marginal statistically significant difference across List 2 Type conditions, $F(2,78)= 2.808$, $p= .066$ (see ANOVA summary Table 2). Recall was higher for the future A-D list condition (M=6.731, SD=5.258) than for the future C-D list condition (M=3.731, SD=3.715), d=.66, p=.062. There was no statistically significant difference in initial recall for the future repeated A-B condition (M=5.035, SD=4.625) list the

two other future lists. The marginally significant effects are unexpected as the manipulation of List 2 type did not occur until after List 1 recall.

An ANOVA (see the summary in Table 3) was conducted on mean correct recall scores for List 2 (see Figure 7). There was a significant effect of task following List 1, $F(2, 232)=3.336$, $p=.037$. Post-hoc tests showed that the JOL task ($M= 11.134$, $SD= 6.427$) produced greater recall than both the cued-recall condition ($M= 9.148$, $SD= 6.193$, $d=.31$, $p=.078$) and restudy condition ($M= 9.038$, $SD= 6.425$, $d=.33$, $p=.082$). There was a significant effect of List 2 type, $F(2, 232)=16.159$, $p<.000$. Post-hoc tests indicated that recall for list two items was significantly higher in the repeated A-B list ($M= 12.723$, $SD= 6.660$) than the C-D list ($M= 8.923$, $SD= 6.239$, $d=.59$, $p< .000$) and the A-D list ($M= 7.588$, $SD= 5.061$, $d=.87$, $p< .000$). The latter two conditions did not differ significantly, $p= .482$. The interaction between task following List 1 and List 2 type was not statistically significant, $F(4, 232)=1.249$, $p= .291$.

In Experiment 1, participants were not required to make a response to every cue in the recall test for List 2. To test whether the total number of words recalled for the second list was affected by the type of review task and list type, a 3 X 3 (Review [cued-recall, JOL, restudy] x List 2 [A-B, A-D, or C-D]) random groups ANOVA was run on the total number of words recalled for List 2 items (see Figure 8; see ANOVA summary Table 4). There was a significant main effect of task following List 1, $F(2,232)=3.423$, $p=.034$. There were significantly more responses following the JOL task ($M=15.354$, $SD=5.444$) than the cued-recall task ($M=13.000$, $SD=5.846$, $d= .42$, $p=.029$). The restudy ($M=14.256$, $SD=6.146$) task did not lead to a different number of words recalled compared to the JOL ($p=.480$) and cued-recall ($p=.723$) tasks. There was a significant main effect of List 2 type, $F(2, 232)=3.742$, $p=.025$. The repeated A-B list ($M=15.566$, $SD=5.575$) had significantly more words recalled than did the C-D list ($M=13.205$,

SD=6.097, *d*=.41, *p*=.029). The repeated A-B list trended towards producing greater recall than the A-D list (*M*=13.775, *SD*=5.750, *p*=.141). There was no difference between the A-D and C-D lists. The differences in the number of responses made on List 2 may be due to the mandatory responses required by JOL for each item that were not mandated for List 1 cued-recall. These differences in overall recall of List 2 for tasks and list type may reflect a response bias. To control for the possible bias, correct recall on the List 2 items was conditionalized on the total number of words produced. That is, the total number of correct responses was divided by the total number of responses made to yield a proportion correct recall score.

A 3 X 3 (Review [cued-recall, JOL, restudy] x List 2 [A-B, A-D, or C-D]) random groups ANOVA was run on the proportion of correct words recalled for List 2 (see Figure 9; see ANOVA summary Table 5). There was no main effect of task following List 1, $F(2, 232)=1.074$, $p= .343$. The main effect of List 2 type was significant $F(2, 232)= 16.742$, $p< .000$. Post-hoc tests showed that the proportion of words correctly recalled was significantly higher for the repeated A-B list (*M*= .806, *SD*= .270) than both the C-D (*M*= .662, *SD*= .325, *d*=.48, *p*= .005) and A-D lists (*M*= .538, *SD*= .262, *d*=1.01, *p*<.000). The proportion of words recalled was significantly higher for the C-D list than the A-D list (*d*=.42, *p*= .037). The interaction between List 2 type and review task following list one approached statistical significance $F(4, 232)= 2.314$, $p= .058$.

Separate one-way ANOVAs tested the simple effect of List 2 (A-B, A-D, C-D) for each type of review task. For the cued-recall condition there was a significant effect of List 2 type, $F(2,80)=7.777$, *p*=.001. The proportion of words correctly recalled was significantly higher in the repeated A-B list (*M*=.847, *SD*=.233) than both the C-D list (*M*=.557, *SD*=.350, *d*=.99, *p*=.016) and the A-D list (*M*=.624, *SD*=.271, *d*=.89, *p*=.001). There was no difference between

the A-D and C-D lists. In the JOL condition there was a significant effect of List 2 type, $F(2,81)=6.901$, $p=.002$. The A-D list ($M=.554$, $SD=.250$) produced significantly lower recall than both the C-D list ($M=.748$, $SD=.270$, $d=.47$, $p=.025$) and the repeated A-B list ($M=.804$, $SD=.270$, $d=.96$, $p=.002$). There was no statistically significant difference between the C-D and repeated A-B lists. Finally, in the restudy condition there was a significant effect of List 2 type, $F(2,77)=6.810$, $p=.002$. The A-D list ($M=.466$, $SD=.250$) produced significantly lower recall than both the C-D list ($M=.681$, $SD=.333$, $d=.73$, $p=.035$) and the repeated A-B list ($M=.763$, $SD=.310$, $d=1.06$, $p=.002$). There was no statistically significant difference between the C-D and repeated A-B lists. This pattern of simple effects suggests that the cued-recall task following the first list eliminates the difference between C-D and A-D items. This is the pattern expected if testing reduces interference.

Separate one-way ANOVAs tested the simple effect of Task (cued-recall, JOL, restudy) for each List 2 type. For the C-D list there was no significant effect of task, $F(2,75)=2.370$, $p=.100$. There was a trend for JOL ($M=.748$, $SD=.270$) to produce greater recall than cued-recall ($M=.557$, $SD=.350$, $p=.104$). There was no difference for restudy ($M=.681$, $SD=.333$) compared to cued-recall or JOL. In the A-D list there was no statistically significant effect of task, $F(2,77)=2.477$, $p=.091$. There was a trend for cued-recall ($M=.624$, $SD=.271$) to produce greater recall than restudy ($M=.466$, $SD=.250$, $d=.61$, $p=.088$). There was no difference between JOL ($M=.554$, $SD=.250$) and cued-recall or restudy. In the repeated A-B list there was no statistically significant effect of task, $F(2,80)=.661$, $p=.519$ where the mean proportion of correct recall was as follows:. cued-recall ($M=.847$, $SD=.233$); JOL ($M=.804$, $SD=.270$); restudy ($M=.763$, $SD=.310$).

A 3 X 2 (review [cued-recall, JOL, restudy] X List 2 [A-B, A-D, C-D]) random groups ANOVA was conducted on the mean number of intrusions from List 1 on List 2 recall (see Figure 10; see ANOVA summary Table 6). There was a significant main effect of List 2 type, such that the C-D list (*M*=.288, *SD*=.716) produced fewer intrusions than the A-D list (*M*=1.363, *SD*=1.561), $F(1,147)=29.572$, *d*=.88, *p*<.000. The main effect of task, $F(2,147)=2.488$, *p*=.087 was marginally significant. The restudy condition produced more intrusions (*M*=1.160, *SD*=1.186) than both the cued-recall (*M*=.776, *SD*=1.159, *p*=.358) and JOL (*M*=.630, *SD*=1.186, *d*=.45, *p*=.085) conditions. There was no significant interaction between List 2 type and task following List 1, $F(2,147)=.853$, *p*=.428.

Correlations were run between average JOL scores and proportion of correct recall scores on List 2, and the proportion of initially correct recall scores following List 1 and the proportion of correct recall scores of List 2. There was a medium positive correlation (*r*= .405, *p*< .000) between average JOL scores (*M*=41.193, *SD*=19.812) on List 1 items and the proportion of correct recall scores on List 2. There was a strong positive correlation (*r*= .601, *p*< .000) between the proportion of initially correct recall scores of List 1 and the proportion of correct recall scores of List 2. The difference between the correlations was not statistically significant (*z*=.91, *p*=.36). Overt and covert retrieval tests seem to converge on individual differences in memory.

Discussion

Data from the proportion of words correctly recalled, found that recall for List 2 items was best in the repeated A-B list followed by the completely novel C-D list and then the interference A-D list. These results are consistent with the first hypothesis. The intrusion results also suggest that being tested on the A-B list reduces interference in the A-D list compared to restudying. That outcome is also consistent with first hypothesis.

As expected, there was no evidence of a testing effect in the recall of the repeated A-B list when looking at the proportion of words correctly recalled. The testing effect is less likely to emerge at a short retention interval (Chan et al., 2018). However, contrary to expectations, there was no statistically significant evidence of a forward testing effect. That is, for the A-D and C-D lists neither cued-recall or JOL produced significantly higher recall than restudy. However, there was a trend for testing both covertly and overtly to produce a forward testing effect. The degree of testing effect and forward testing effects can be determined by calculating the difference between the mean proportion of words recalled for either the cued-recall or JOL tasks to the restudy task. JOL recall levels were 10% higher than restudy recall levels for the A-D list and 7% higher for the C-D list. Cued-recall recall levels were 18% higher than restudy recall levels for the A-D list. Within the C-D list, there was not a forward testing effect found for the cued-recall condition. The trend for A-D to produce larger forward testing effects than C-D supports the second hypothesis. The trends are also higher for overt than covert retrieval.

The second hypothesis was further supported by the greater number of intrusions that occurred in the A-D list compared to the C-D list, and the trend for testing to reduce intrusions in the A-D list. The highest number of intrusions occurred in the restudy condition. These results support the theory that testing reduces proactive interference (Tulving & Watkins, 1974) and are consistent with integration theory. That is, testing leads to a binding of old and new targets to a common cue.

The third hypothesis centred on TAP as a mediator of the testing effect and the forward testing effect such that overt retrieval (cued-recall) should lead to stronger effects than covert retrieval (JOL). The results of the first experiment trended towards partial support for this hypothesis. The repeated A-B list had a greater testing effect with the cued-recall task (8%) than

with the JOL task (5%). The A-D list had a greater forward testing effect for the cued-recall task (18%) than the JOL task (11%). For the C-D list the JOL task trended towards producing a forward testing effect (7%), but cued-recall led to lower recall levels than restudy. The low level of recall on the C-D list may be attributed to the low level of recall that occurred on the List 1 recall of the future C-D list condition.

Figure 7, depicting the mean number of correct responses, shows that, numerically, the JOL task produced a testing effect (repeated A-B list) and forward testing effects (A-D, and C-D). Overt retrieval appears to be the same as covert retrieval in yielding a trend of a forward testing effect for the A-D list. Although the mean number of words recalled was higher for the JOL conditions, these results could have derived from the higher levels of attempted retrieval compared to cued-recall. For this reason, the proportion of correct words recalled (i.e., what proportion of words produced are correct) may be a more sensitive index of the benefits of testing. There appears to be partial support for TAP, with overt retrieval producing larger effects than covert retrieval in the A-D and repeated A-B lists. Overall, the results are consistent with past research which has found that both overt and covert retrieval lead to a testing effect and a forward testing effect (Cho et al., 2017; Sundqvist et al., 2017).

In conclusion, the results of the experiment appear to in part be due to TAP, as indicated by the trend for cued-recall to produce a larger testing effect and forward testing effects for than JOL in the A-D and repeated A-B lists. The results supported past research showing that being tested reduced proactive interference. The strongest evidence in that regard is the lower levels of intrusions in the A-D list in the overt and covert conditions than in the restudy condition. The results of the experiment appear to support integration theory which predicts that the A-D list would produce a larger forward testing effect than C-D list because previously being tested for

List 1 items spontaneously prompts the previous target that was paired with the cue when an individual learns a new list containing part of the original list (i.e. the cue word).

Differences between recall of List 2 following JOL and cued-recall conditions may be due to the requirement that JOL participants respond to each item. For Experiment 2, instructions for the cued-recall task following the first list required participants to respond to each item to reduce a potential bias effect that may have led to fewer overall responses to List 2 items. Experiment 2 used a within-subjects design for List 2 item type as in Cho et al. (2017) and Wahlheim (2015). Mixing repeated A-D and C-D items in the second list may work against integration of responses across lists so as to diminish forward testing effects.

Experiment 2 (Mixed Design: List Type as a Repeated Measure)

Method

Material. I used the word-pairs from Wahlheim (2015), but to stay consistent with Tulving and Watkins (1974) I re-paired the items so that they would be unrelated. The Wahlheim (2015) word-pairs were derived from the University of Florida word association, rhyme, and word fragment norms data base (Nelson, McEvoy & Schreiber, 1998). The word-pair lists were increased from 24 word-pairs to 30 word-pairs so that the second list contained 10 pairs comprising each of the A-B, 10 A-D, and 10 C-D conditions (see Appendix A). A second version of these lists was created as in the first experiment. The word-pair presentation length was increased from four seconds to six seconds with a one second interstimulus interval to increase the recall level. A new consent and debriefing form were created for this experiment (see Appendix D for the forms titled Metamemory and Learning).

Procedure. A total of 79 participants was recruited (27 overt retrieval; 26 covert retrieval; 26 restudy). Participants first viewed 30 A-B word-pairs following which they

completed one of three tasks: (1) overt retrieval; (2) covert retrieval; or, (3) restudy. All groups then studied a mixed A-B/A-D/C-D list, followed by a cued-recall test of that mixed list.

The second experiment followed the same procedure as Experiment 1 with the exception that participants performing the cued-recall task were also instructed to respond to every item. They were asked to guess if they were unable to recall the missing word. The first and second lists each contained 30 word-pairs. In the second list, 10 of the word-pairs were from the repeated A-B items, 10 were A-D items, and 10 were C-D items. The 'D' items in the A-D and C-D conditions were unique within the list. Following presentation of List 1, participants completed either a restudy, cued-recall, or JOL task before viewing the second list. Following the presentation of the second list all participants completed a cued-recall task for that list.

Analysis. A 3 X 3 (Review [cued-recall, JOL, restudy]) x List 2 item [A-B, A-D, or C-D]) mixed ANOVA was run on recall scores of List 2. Review type was varied between-subjects and item type was varied within-subjects. A 3 X 2 (Review [cued-recall, JOL, restudy]) x List 2 item [A-D or C-D]) mixed ANOVA was conducted on the number of intrusions that occurred in the second list. A correlational analysis was run between initial number of items correctly recalled following List 1 and List 2 recall scores, as well as between the average JOL and final recall scores. These correlations were run for the total items recalled, and for each item type. For all analyses an alpha level of 0.05 was used, and a Bonferroni correction was applied for any post-hoc comparisons.

Results

A 3 X 3 (Review [cued-recall, JOL, restudy] x List 2 Item [A-B, A-D, or C-D]) mixed ANOVA with item as a repeated-measure was run on List 2 recall scores. The ANOVA showed that there was no significant main effect of task following the first list, $F(2, 76)= .003$, $p= .997$

(see Figure 11; see ANOVA summary Table 7.). There was a significant main effect of List 2 items, $F(2, 152)= 29.547$, $p< .000$. Pairwise comparisons showed that the repeated A-B items ($M= 4.443$, $SD=2.581$) produced significantly greater recall than both the A-D items ($M= 2.734$, $SD= 2.086$, $d=.75$, $p< .000$) and the C-D items ($M= 3.165$, $SD=2.420$, $d=.57$, $p< .000$). There was no difference between A-D and C-D items. The interaction between Review and List 2 approached statistical significance, $F(4,152)= 2.301$, $p= .061$.

To explore the source of that interaction, simple effects of Review Task were tested for each Item type. For the repeated A-B items there was no effect of task, $F(2,76)=.667$, $p=.516$. Recall was equivalent for cued-recall ($M= 4.185$, $SD= 2.370$), JOL ($M= 4.185$, $SD= 4.231$), and restudy ($M= 4.923$, $SD= 2.938$). Although the A-D items showed that recall was lowest for the restudy group, the simple effect of task for the A-D items was not statistically significant, $F(2,76)=.912$, $p=.406$. Recall was statistically equivalent for cued-recall ($M= 2.815$, $SD= 2.131$), JOL ($M= 3.077$, $SD= 2.171$), and restudy ($M= 2.308$, $SD= 1.955$). The C-D items also showed no statistically significant simple effect of task, $F(2,76)=.228$, $p=.797$. Recall was equivalent for cued-recall ($M= 3.407$, $SD= 2.135$), JOL ($M= 2.961$, $SD= 2.391$), and restudy ($M= 3.115$, $SD= 2.776$). There was a trend for the JOL and cued-recall to produce a forward testing effect for the A-D items. However, for the C-D items JOL recall was lower than for restudying and cued-recall. By having participants respond to each item the differences between the overt (cued-recall) and covert (JOL) retrieval methods were eliminated. This supports the interpretation of JOL superiority in Experiment 1 as due to having to make responses for each item during the JOL condition but not the cued-recall condition.

To further probe the source of the trend of a Review Task X Item interaction, the simple effect of Item was tested for each Review Task. For the cued-recall condition there was a

significant effect of List 2 items $F(2, 52)= 4.506$, $p= .016$. There was significantly higher recall for the repeated A-B items ($M= 4.185$, $SD= 2.370$) than A-D items ($M= 2.815$, $SD= 2.131$, $d=.52$, $p= .036$). Recall was higher for the repeated A-B items than C-D items ($M= 3.407$, $SD= 2.135$) but the difference was not statistically significant ($p=.236$). Recall of the C-D items was greater than A-D items, but the difference was not statistically significant ($p=.562$).

For the JOL group there was a significant effect of List 2 item, $F(2, 50)= 7.245$, $p= .002$. Recall of the repeated A-B items ($M= 4.231$, $SD= 2.438$) was significantly greater than the A-D ($M= 3.077$, $SD= 2.171$, $d=.59$, $p= .019$) and C-D items ($M= 2.961$, $SD= 2.391$, $d=.67$, $p= .006$). There was no statistically significant difference between the A-D and C-D items. Finally, the restudy group showed a statistically significant effect of List 2 items, $F(2, 50)= 26.672$, $p< .000$. The repeated A-B items ($M= 4.923$, $SD= 2.938$) had significantly higher recall than the A-D items ($M= 2.308$, $SD= 1.955$, $d=1.31$, $p< .000$), and the C-D items ($M= 3.115$, $SD= 2.776$, $d=1.17$, $p< .000$). Although the C-D items produced greater recall than the A-D items, the difference was not statistically significant ($p= .157$) Overall the results suggest that recall was higher for the repeated items than novel items (A-D and C-D).

Interference of List 1 on List 2 learning was analyzed for A-D and C-D items with a 3 X 2 (Review [cued-recall, JOL, restudy] X List 2 Item [A-D, C-D] mixed ANOVA. The total number of intrusions varied across tasks $F(2, 76)= 14.511$, $p< .000$ (see Figure 12; see ANOVA summary Table 8). The restudy condition produced significantly more intrusions ($M= 2.192$, $SD= .849$) than both the cued-recall ($M= .519$, $SD= .849$, $d=1.97$, $p< .000$) and JOL ($M= .692$, $SD= .884$, $d=1.73$, $p< .000$) conditions. There was no difference between JOL and cued-recall ($p=1.000$). There was a main effect of List 2 item type $F(1,76)=49.096$, $p<.000$. The A-D items ($M= 1.025$, $SD= 1.349$) produced significantly more intrusions than the C-D items ($M= .101$,

SD= .343), *d*=.69, *p*< .000. There was a significant interaction between Review Task and List 2 Item, *F*(2,76)=12.847, *p*<.000.

The simple effect of Item was tested for each task. For the cued-recall task A-D items (*M*=.482, *SD*=.849) produced significantly more intrusions than C-D items (*M*=.370, *SD*=.192, *d*=.18, *p*= .016). For the restudy task A-D items (*M*=2.039, *SD*=1.661) produced significantly more intrusions than C-D items (*M*=.154 , *SD*=.464, *d*=1.546, *p*< .000). Finally, for the JOL task A-D items (*M*=.577, *SD*=.758) produced significantly more intrusions than C-D items (*M*=.115, *SD*=.326, *d*=.79, *p*= .005). The simple effect of Review Task was tested for both the A-D and C-D items. For the A-D items there was a significant effect of task, *F*(2, 76)= 14.858, *p*< .000. Restudy (*M*= 2.039, *SD*= 1.661) produced significantly more intrusions than both cued-recall (*M*= .485, *SD*= .849, *d*=1.18, *p*< .000) and JOL (*M*= .577, *SD*= .758, *d*=1.14, *p*< .000). There was no difference between JOL and cued-recall (*p*=1.000). For C-D items there was no significant difference across tasks, *F*(2, 76)= .746, *p*=.455. There was a trend for the restudy condition to produce more intrusions (*M*= .154, *SD*= .464) than, JOL (*M*= .115, *SD*= .326), and cued-recall (*M*= .037, *SD*= .192).

Item-specific intrusions for A-D items in List 2 (recalling the target that was previously paired with the cue word in the A-B list) were analysed. There was a significant effect of task on item-specific intrusions, *F*(2,76)= 17.262, *p*<.000. Restudy led to significantly more item-specific intrusions (*M*= 1.692, *SD*= 1.619) than both the cued-recall (*M*= .296, *SD*= .465, *d*=1.17, *p*<.000) and JOL (*M*= .269, *SD*= .452, *d*=1.21, *p*<.000). There was no difference between the number of item-specific intrusions occurring in the JOL and cued-recall conditions (*p*=1.000). These results suggest that both overt and covert retrieval methods reduce item-specific interference.

The correlation between mean JOL scores and overall Final Recall (recall of the three types of list items in List 2 combined) was large (r= .631, p= .001). The large positive correlation between JOL scores of list one and final recall held for each of the three item types: repeated A-B items (r=.545, p= .004); A-D items (r= .486, p= .012); and, C-D items (r= .642, p< .000);. The correlation between initial recall of List 1 and overall final recall of List 2 was statistically significant (r= .687, p< .000). The initial correct recall scores was then correlated with the correct recall within each List 2 item type. All items yielded significant large positive correlations: repeated A-B items (r= .597, p =.001); A-D items (r= .446, p=.020); and, a C-D items (r= .573, p= .002). Overt and covert retrieval tests converge on individual differences in memory.

Discussion

The first hypothesis predicted greater recall for the repeated A-B items, followed by C-D items and then A-D items. The results of Experiment 2 found partial support for the hypothesis in that the repeated A-B items produced significantly greater recall than both the A-D and C-D items. However, the A-D and C-D items did not differ.

The second hypothesis correctly predicted that there would be no testing effect. The second hypothesis also predicted that the A-D items would produce a larger forward testing effect than the C-D item. Although no statistically significant forward testing effects were found, the A-D list trended towards producing a larger forward testing effect than C-D items. This can be seen in the difference in the magnitude of the forward testing effects, with the cued-recall condition producing a larger effect for the A-D items (5%) than the C-D items (3%). Within the JOL condition, the A-D list trended towards producing a forward testing effect (8%), while there was no difference in the C-D list (2%). Those patterns are consistent with the second hypothesis.

Further support for the second hypothesis is found in the intrusion data which showed that there were significantly more intrusions for A-D than C-D items. There data found that both overt and covert retrieval significantly reduced the number of overall and item-specific intrusions compared to restudy. The results also found that this effect was driven by the A-D list, with both forms of testing significantly reducing both intrusions and item specific intrusion compared to restudy. This further supports the integration theory for the forward testing effect. This is the same pattern found in Experiment 1.

The third hypothesis predicted that overt retrieval would produce a larger testing effect and larger forward testing effects than covert retrieval. Since no significant testing effect or forward testing effects were found, there is not support for TAP. The results found that there was no difference in recall levels between overt and covert retrieval, although both trended towards producing forward testing effects. This is consistent with Experiment 1. This is consistent with the findings of Sundqvist et al. (2017) that overt and covert retrieval produce equivalent benefits.

Overall, both overt (cued-recall) and covert (JOL) retrieval methods trend towards producing forward testing effects. The differences between overt and covert retrieval from Experiment 2 suggested that the results shown in the Experiment 1, with JOL producing greater levels of retrieval, were likely due to the nature of responding to all items following list one when participating in the JOL condition. Similar to Experiment 1, the completely novel (C-D) items produced greater recall than the interference (A-D) items. The results also suggest that being tested on the A-B list reduces interference in the A-D items compared to restudying, as shown by the reduced number of intrusions.

General Discussion

The seminal study of the forward testing effect by Tulving and Watkins (1974) led to the proposal that testing on the initial A-B list reduced proactive interference when learning an A-D list. However, they did not include a control C-D list or a restudy baseline. Wahlheim (2015) compared restudy and cued-recall with a within-subjects design. The relationship of the second list items to the first list items (repeated A-B item, interference A-D items and C-D items) was also varied within-subjects. Neither of these studies examined the role of covert retrieval on the testing effect and the forward testing effect. The present study sought to fill these gaps. The review task was varied between-subjects (as in Tulving &Watkins, 1974). The comparison of covert and overt retrieval methods was designed to test the TAP explanation of the benefits of retrieval. Across two experiments, item type was varied between-subjects in Experiment 1 and within-subjects in Experiment 2.

The first hypothesis predicted that recall would be highest for the repeated A-B items, followed by the C-D items and then the A-D items. Across both experiments, the repeated A-B items were recalled at a significantly higher rate than the C-D items and A-D items. Both experiments found that there were no differences in the recall of the A-D and C-D items. It was predicted that the C-D items would have greater recall than the A-D items due to the proactive interference in the A-D items restudy condition. Both experiments found that the A-D list produced the greatest number of intrusions in the A-D restudy condition. However, the number of intrusions was small for both experiments which may explain why there was no difference between the A-D and C-D items.

The second hypothesis, that there would be no testing effect, was supported by both experiments. Neither experiment found a significant forward testing effect. However, there were

trends for the A-D list to produce larger forward testing effects than C-D items as predicted by integration theory.

The second hypothesis predicted that the larger forward testing effect for the A-D items was due to a reduction in intrusions. The results from both experiments confirmed this hypothesis. A-D items produced more intrusions than the C-D items. The restudy condition led to more intrusions than did both the overt and covert retrieval conditions. There was a trend for cued-recall to produce the lowest number of intrusions. That pattern supports integration theory which assumes that testing on the initial list spontaneously cues an individual to recall the first target and the new target when they are encoding information in a novel list (Chan et al., 2018). Integration can only occur when there is some relationship between targets as was the case for the A-D items which shared stimuli with the original A-B list. Integration may be the mechanism by which testing mitigates proactive interference and reduces intrusions from the first list (Tulving & Watkins, 1974).

Experiment 1 showed a trend, in the proportion of words correctly recalled for List 2, for covert retrieval to produce a general benefit where all lists saw improved recall compared to restudying. Overt retrieval showed a trend for an item specific benefit, where overt retrieval improves recall of A-D and repeated A-B lists. However, it impairs recall of the C-D list. This result was unexpected and likely due to the low levels of initial recall in future C-D list condition.

The third hypothesis predicted that overt retrieval would produce a larger testing effect and forward testing effect than covert retrieval under TAP. The first experiment found partial support for TAP in the repeated A-B and A-D list where there was a trend for cued-recall to produce a larger testing effect and forward testing effect than JOL. The second experiment also

found a trend for cued-recall to produce a forward testing effect in the C-D list. The reason that cued-recall in the first list did not produce a larger effect than JOL may be due to the initial low levels of items recalled during List 1 in the cued-recall condition. Overall, these results support past research which has found that both overt and covert retrieval are effective in generating the testing effect and forward testing effect (Cho et al., 2017; Sundqvist et al., 2017).

The results from Experiment 1 suggest that covert retrieval is more effective than overt retrieval when the mean number of correct responses was analyzed. However, when response rates were equated in the second experiment, by mandating cued-recall responses for each item, the difference between JOL and cued-recall was not statistically significant.

TAP theory does not fully account for the differences between overt and covert retrieval since the results from both experiments did not reach statistical significance. In Experiment 1 the trend for overt retrieval to impair the learning of the C-D list goes against TAP, additionally if just the mean number of responses are analyzed there were higher recall scores for JOL than cued-recall. TAP is able to partly account for why there were trends in both experiments for overt retrieval to produce larger effects than covert retrieval. It is possible that one theory alone may not account for the effects of the testing effect or forward testing effect. One explanation is metacognitive theory where testing either overtly or covertly changes the individual's retrieval or encoding strategies. It assumes that testing draws the individual's attention to how much information they do not know, and this causes the individual to change their retrieval strategy (Chan et al., 2018). If both overt and covert testing both draw the individual's attention to the amount of information they do not know, it may account for why overt and covert retrieval levels were similar. It also appears that integration theory accounts for the trends found in the A-D items.

Although there were several important trends in the results, recall differences relevant to the testing effect and the forward testing effect failed to reach statistical significance. A restudy condition was used which has been shown to produces weaker effects than other review tasks (Chan et al., 2018). The restudy condition provides higher levels of recall to compare testing to and this makes it more challenging to produce a testing effect and forward testing effect. The mixed design may not have found significant forward testing effects since the forward testing effect have been shown to be weaker in within-subject designs (e.g., Wahlheim, 2015, but in his study both item and task are manipulated within-subjects). Having a mixed List 2 may make it more challenging for individuals to try to integrate the two lists together. The restudy baseline used to measure testing effects and the design of Experiment 2 are likely the reasons that the results of the experiments did not reach significance. A larger sample size may also be needed as demonstrated by the trend for the effect sizes to indicate a larger testing effect for A-D than C-D items, and more interference for the restudy than cued-recall conditions for both experiments. This suggests that the sample sizes derived from Tulving and Watkins (1974) were too small.

Another reason that the results of the experiments may not have replicated the forward testing effects in the A-D items as found in past literature may be due to change detection. Wahlheim (2015) found that recall rates were higher when change detection was occurring. In the present experiments, change detection was not assessed. It is possible that the results of the study are tied to the participants' abilities to detect change from the first list to the second list. Future studies should include change detection tasks to gain a better understanding of the underlying mechanisms of the forward testing effect. That is, if the effects are in part driven by an individual's ability to notice change from one list to the next. If the forward testing effect for the A-D list is driven by change detection, it would be expected that the more items individuals

are able to identify as having changed from the first list, the better their final recall of A-D items will be and they will have fewer intrusions.

In conclusion the results of these experiments appear to support integration theory as is shown through trend for the A-D list to produce greater effects than the C-D list. Integration theory is also supported by the intrusion data which showed in both experiments that testing either overtly or covertly reduced interference.

Future studies may address some of the limitations of the present study by: implementing a filler task in place of the restudy condition to produce a larger testing effect and forward testing effect; keeping the experiment between-subjects; increasing sample sizes; comparing related and unrelated word-pairs to determine if related word-pairs produce larger testing and forward testing effects than unrelated pairs as predicted by integration theory; using a change recollection task to help further identify if the integration theory accounts for the larger forward testing effect in the A-D list. Future studies can test the integration theory by attempting to find scenarios in which the forward testing effect does not occur. Including a distractor task between the review of the first list and the encoding of the second list should diminish the forward testing effect because it would make it more challenging for individuals to integrate the two lists together. Such manipulations may reveal the conditions that optimize the benefits of testing on learning.

References

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445-459.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes, 2,* 35-67. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Buchin, Z. L., & Mulligan, N. W. (2017). The testing effect under divided attention, *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000427

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychological Bulletin, 13*(5), 826-830. https://doi.org/10.3758/BF03194004

Chan, J., Meissner, C., & Davis, S. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*(11), 1111-1146.

Cho, K. W., Neely, J. H., Crocco, S., and Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items, *The Quarterly Journal of Experimental Psychology, 70*(7), 1211-1235. http://dx.doi.org/10.1080/17470218.2016.1175485

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology, 6*, 1-6. doi:10.3389/fpsyg.2015.01054

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Mulligan, N. W., Susser, J. A., & Smith, S. A. (2016). The testing effect is moderated by experimental design. *Journal of Memory and Language, 90*, 49-65. https://doi.org/10.1016/j.jml.2016.03.006.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms [Database]. Retrieved from http://w3.usf.edu/FreeAssociation/

Pastötter, B., & Bäuml, K. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5,* 1- 5. doi:10.3389/fpsyg.2014.00286

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*(10), 1-8. https://doi.org/10.3389/neuro.11.010.2008

Racsmány, M., Szőllősi, Á., & Bencze, D. (2017). Retrieval practice makes procedure from remembering: An automatization account of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication, 1-10. http://dx.doi.org/10.1037/xlm0000423

Rawson, K. A., Wissman, K. T., & Vaughn, K. E. (2015). Does testing impair relational processing? Failed attempts to replicate the negative testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1326-1336. http://dx.doi.org/10.1037/xlm0000127

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review,* 1-23. https://doi.org/10.3758/s13423-017-1298-4

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. https://dx.doi.org/10.1037/a0037559

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert vs. overt retrieval. *Frontiers in Psychology, 8,* 1-15. doi:10.3389/fpsyg.2017.01018

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13*, 181-193.

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory, 23*(8), 1229-1237. http://dx.doi.org/10.1080/09658211.2014.970196

Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory and Cognition*, *43*(1), 27-38. https://doi.org/10.3758/s13421-014-0455-5

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning*, *3*(8). doi:10.1038/s41539-018-0024-y

Tables

Table 1

Experiment Design

Experiment 1: Between-Subjects Design: 9 Groups

	List 1 Review		
	Overt Retrieval	Covert Retrieval	Restudy
List 2 Structure			
A-B	A-B	A-B	A-B
A-D	A-D	A-D	A-D
C-D	C-D	C-D	C-D

Experiment 2: Mixed-Design: 3 Groups

	List 2 Review (Between-Subjects)		
	Overt Retrieval	Covert Retrieval	Restudy
List 2 Item (Within-Subjects):			
A-B	A-B	A-B	A-B
A-D	A-D	A-D	A-D
C-D	C-D	C-D	C-D

Table 2

Experiment 1: ANOVA summary table for the number of words correctly recalled for List 1.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
List	117.72	2	58.86	2.81	0.07	0.07
Error	1635.20	78	20.96			

Note. Type III Sum of Squares

Table 3

Experiment 1: ANOVA summary table for List 2 correct recall scores

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
List	1148.80	2	574.40	16.16	< .00	0.12
Review	237.20	2	118.60	3.34	0.04	0.03
List ✻ Review	177.50	4	44.38	1.25	0.29	0.02
Error	8246.80	232	35.55			

Note. Type III Sum of Squares

Table 4

Experiment 1: ANOVA summary table for the total items recalled in List 2.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
List	248.91	2	124.46	3.74	0.03	0.03
Review	227.67	2	113.84	3.42	0.03	0.03
List ✻ Review	74.97	4	18.74	0.56	0.69	0.01
Error	7716.34	232	33.26			

Note. Type III Sum of Squares

Table 5

Experiment 1: ANOVA summary table for the proportion of items correctly recalled for List 2.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
List	2.69	2	1.34	16.74	< .00	0.13
Review	0.17	2	0.09	1.07	0.34	0.01
List ✻ Review	0.74	4	0.19	2.31	0.06	0.04
Error	18.63	232	0.08			

Note. Type III Sum of Squares

Table 6

Experiment 1: ANOVA summary table for the mean number of intrusions for A-D and C-D lists.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
List	44.09	1	44.09	29.57	< .00	0.17
Review	7.42	2	3.71	2.49	0.09	0.03
List ✻ Review	2.55	2	1.27	0.85	0.43	0.01
Residual	219.15	147	1.49			

Note. Type III Sum of Squares

Table 7

Experiment 2: ANOVA summary table for the mean correct recall of List 2 items.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
Review	0.01	2	0.04	<0.00	1.00	<0.00
Error	973.17	76	12.81			
Item Type	125.56	2	62.78	29.55	<.00	0.28
Item Type ✻ Review	19.56	4	4.89	2.30	0.06	0.06
Error	322.96	152	2.13			

Note. Type III Sum of Squares

Table 8

Experiment 2: ANOVA summary table for mean number of intrusions.

Source	Sum of Squares	df	Mean Square	F	p	η^2_p
Review	22.21	1	50.81	66.40	< .00	0.47
Error	38.71	3	12.90	18.99	< .00	0.28
Item Type	34.17	1	34.17	49.096	< .00	0.39
Item Type ✻ Review	17.88	2	8.94	12.85	< .00	0.25
Error	52.89	76	0.70			

Note. Type III Sum of Squares

Figures

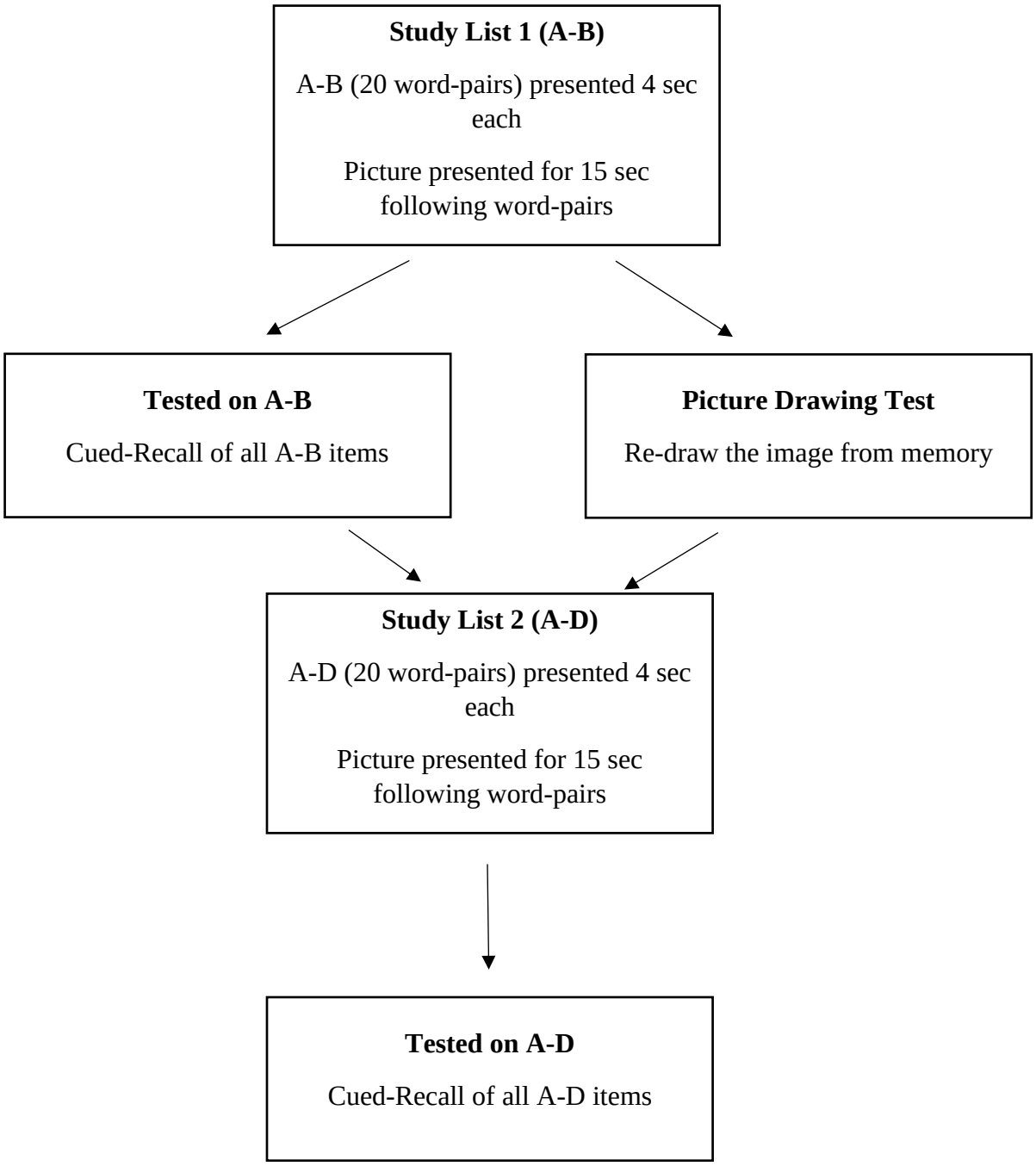

Figure 1. Tulving and Watkins (1974) Procedue

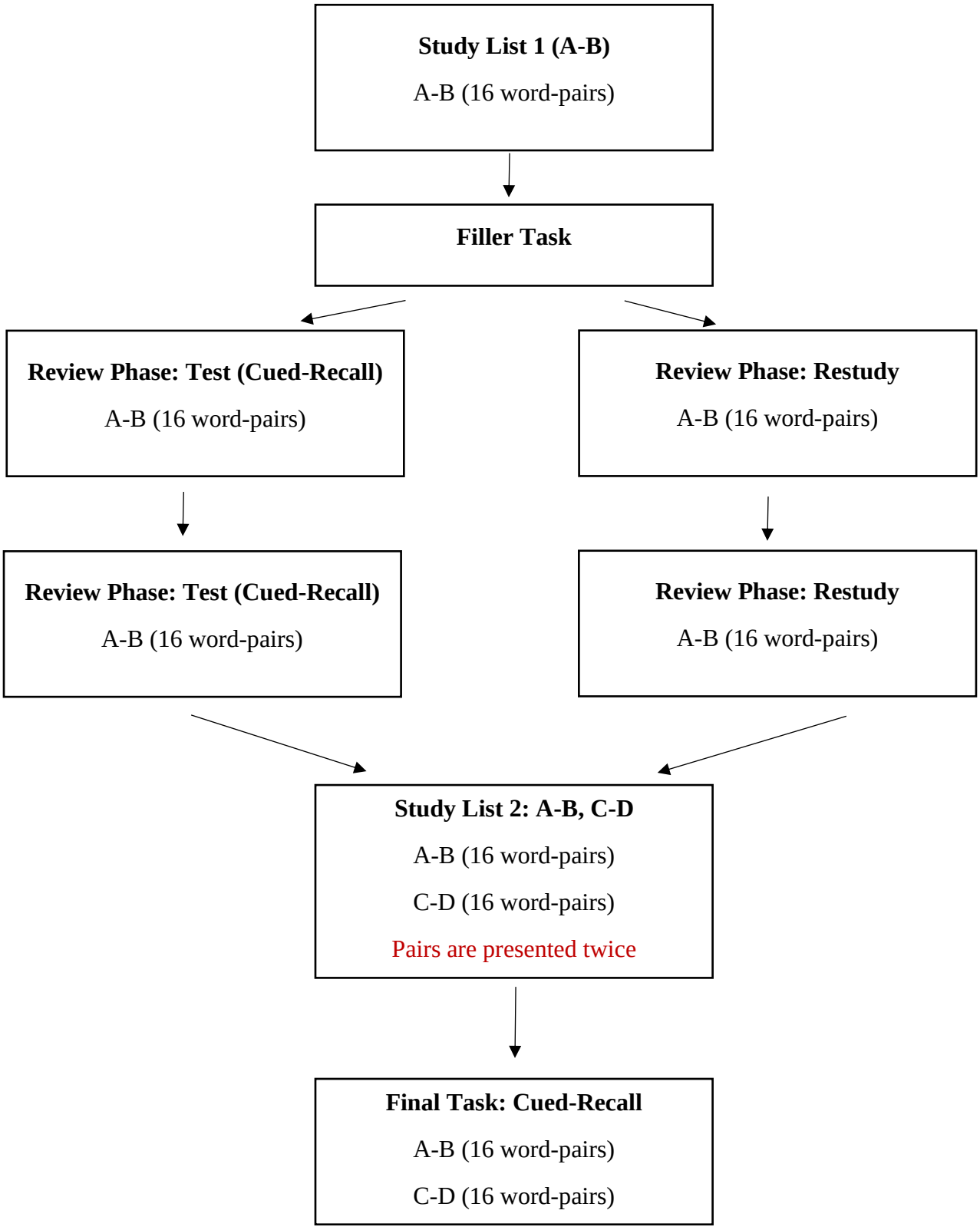

Figure 2. Cho et al. (2017) Procedure

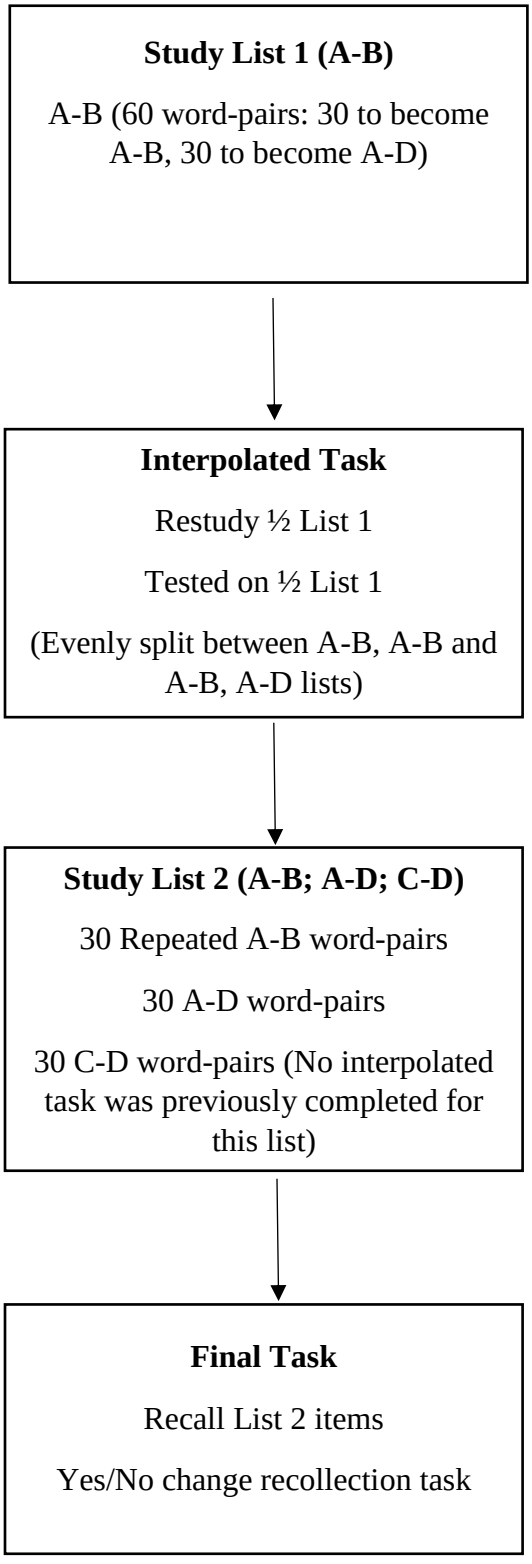

Figure 3. Wahlheim (2015) Procedure

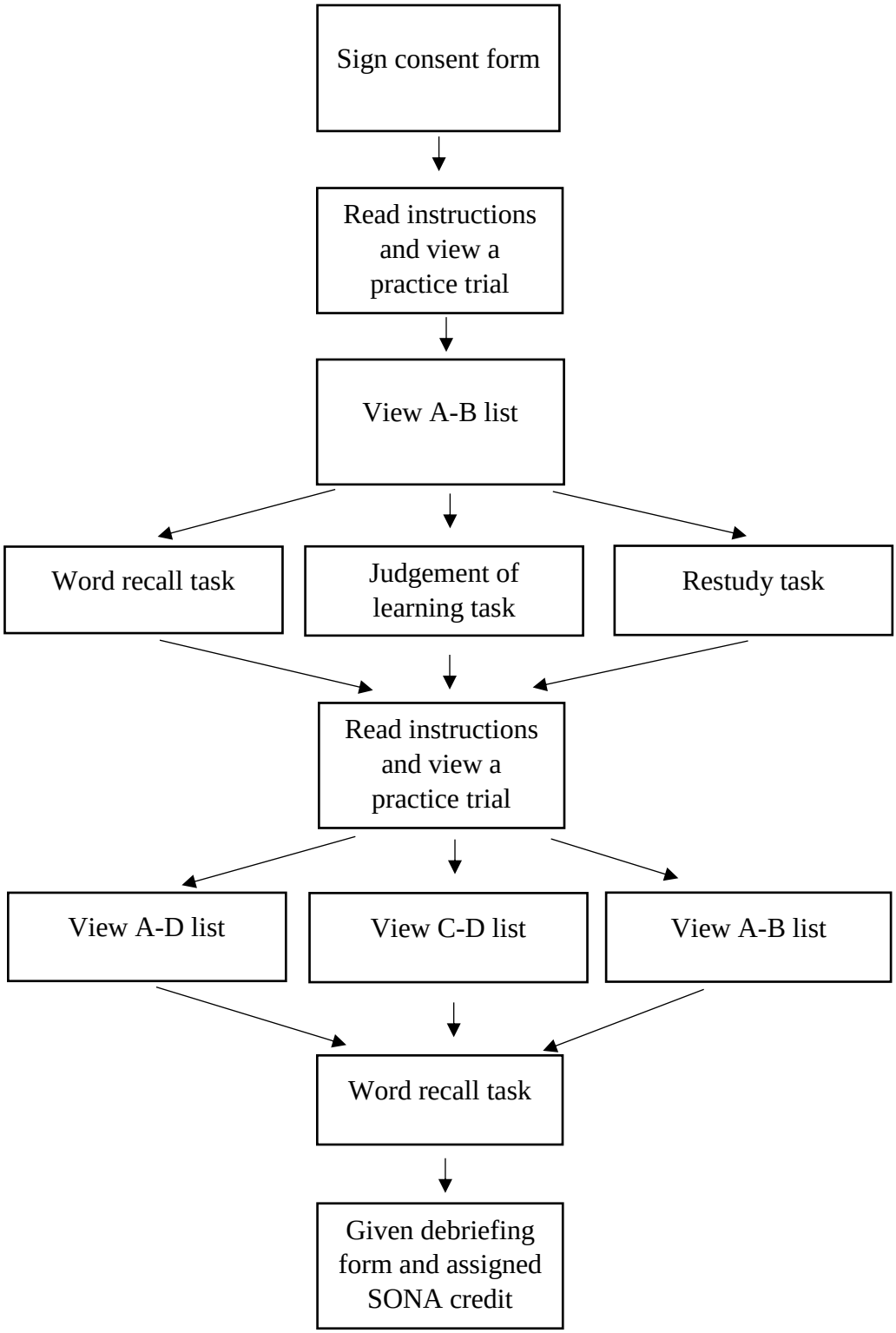

Figure 4. Procedure for Thesis: Between-Subjects Design

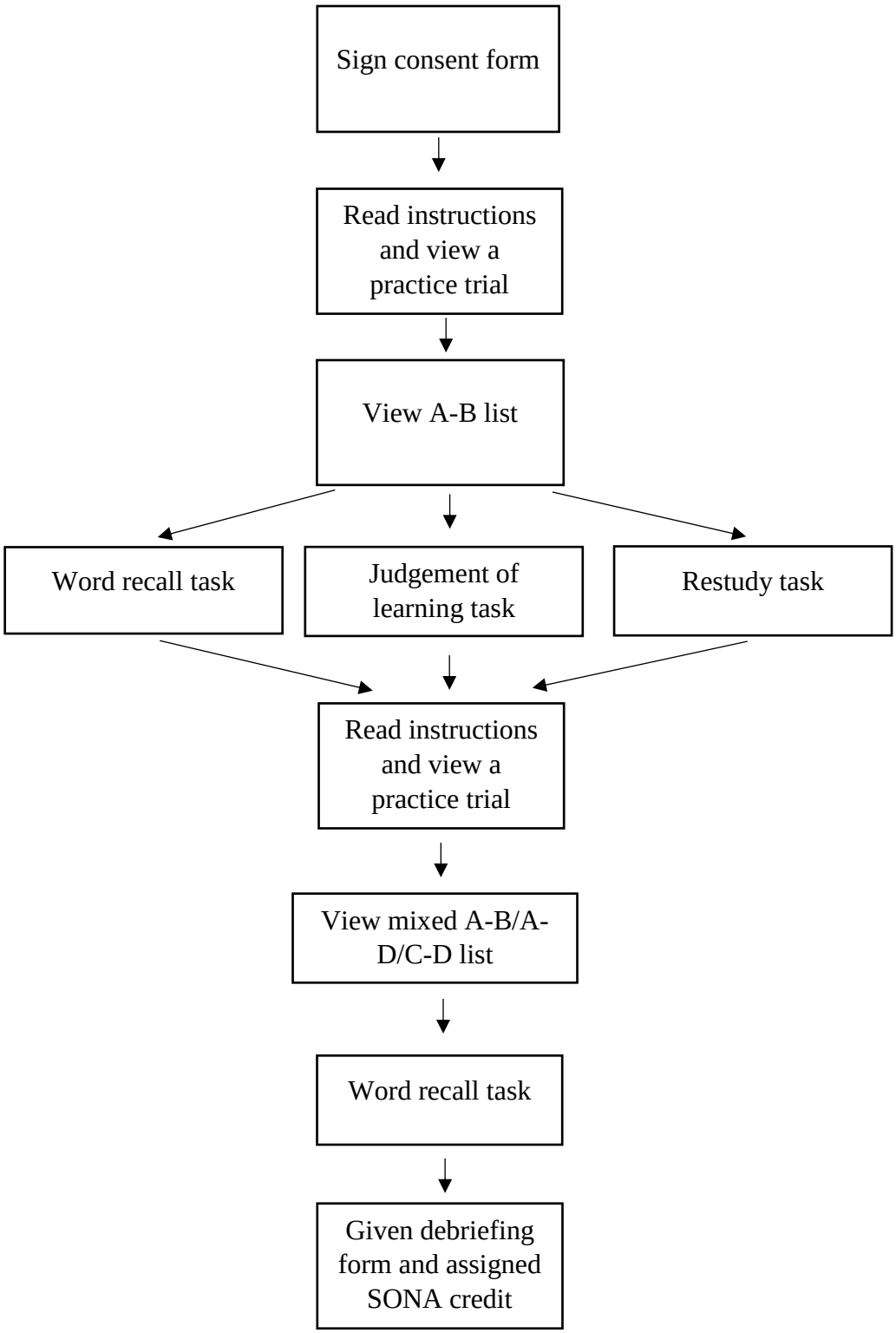

Figure 5. Procedure for Thesis: Mixed-Design

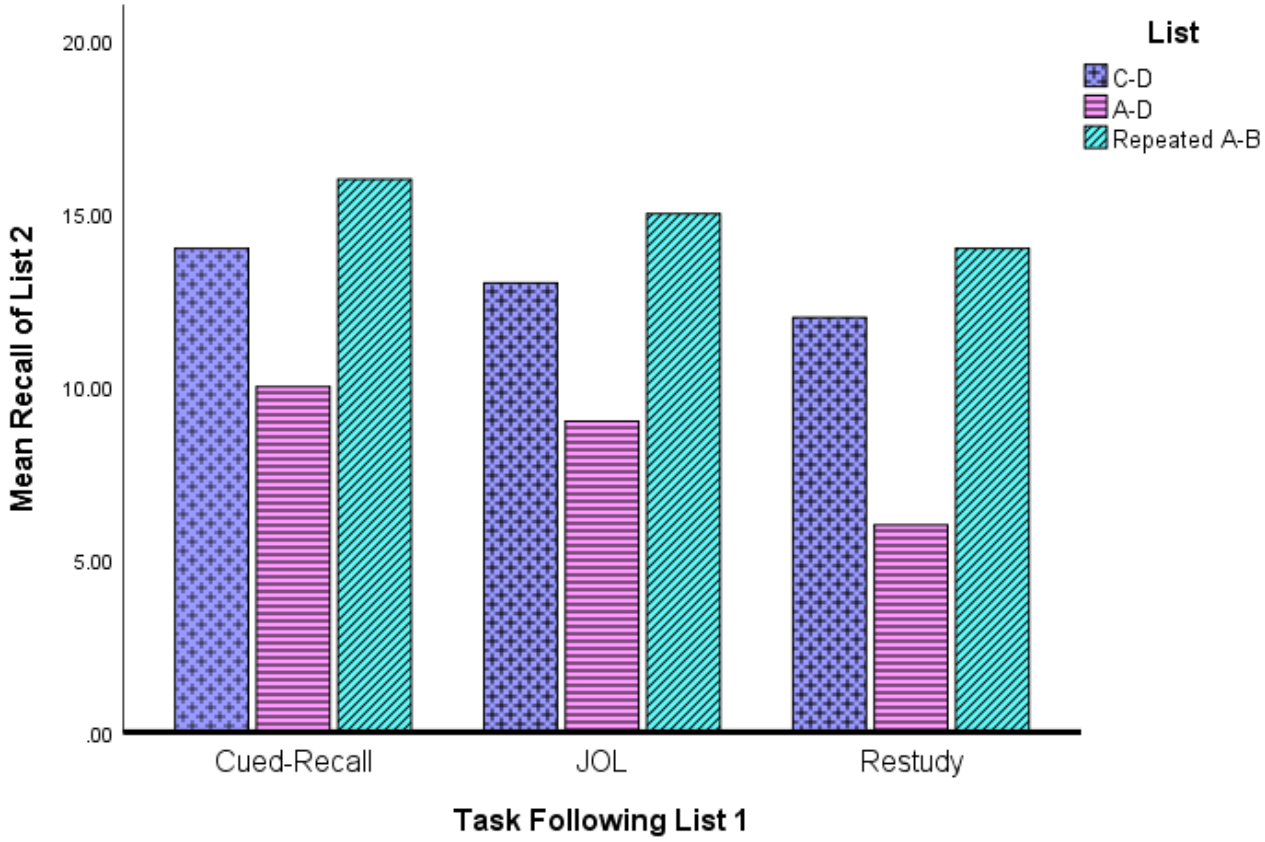

Figure 6. Expected Correct List 2 Recall (Maximum = 24).

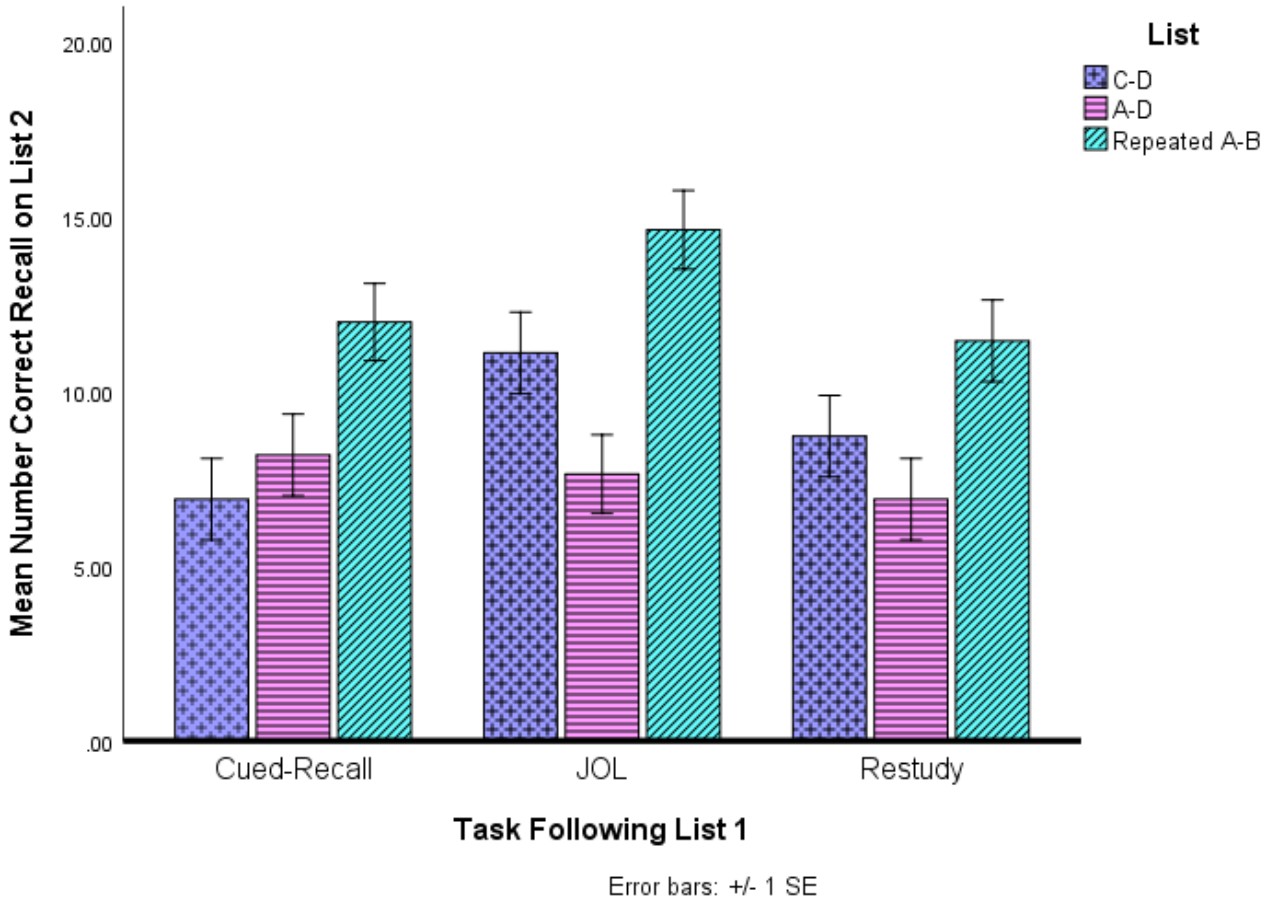

Figure 7. Experiment 1: Mean Number Correct Recall Scores of List 2 Words (Maximum = 24).

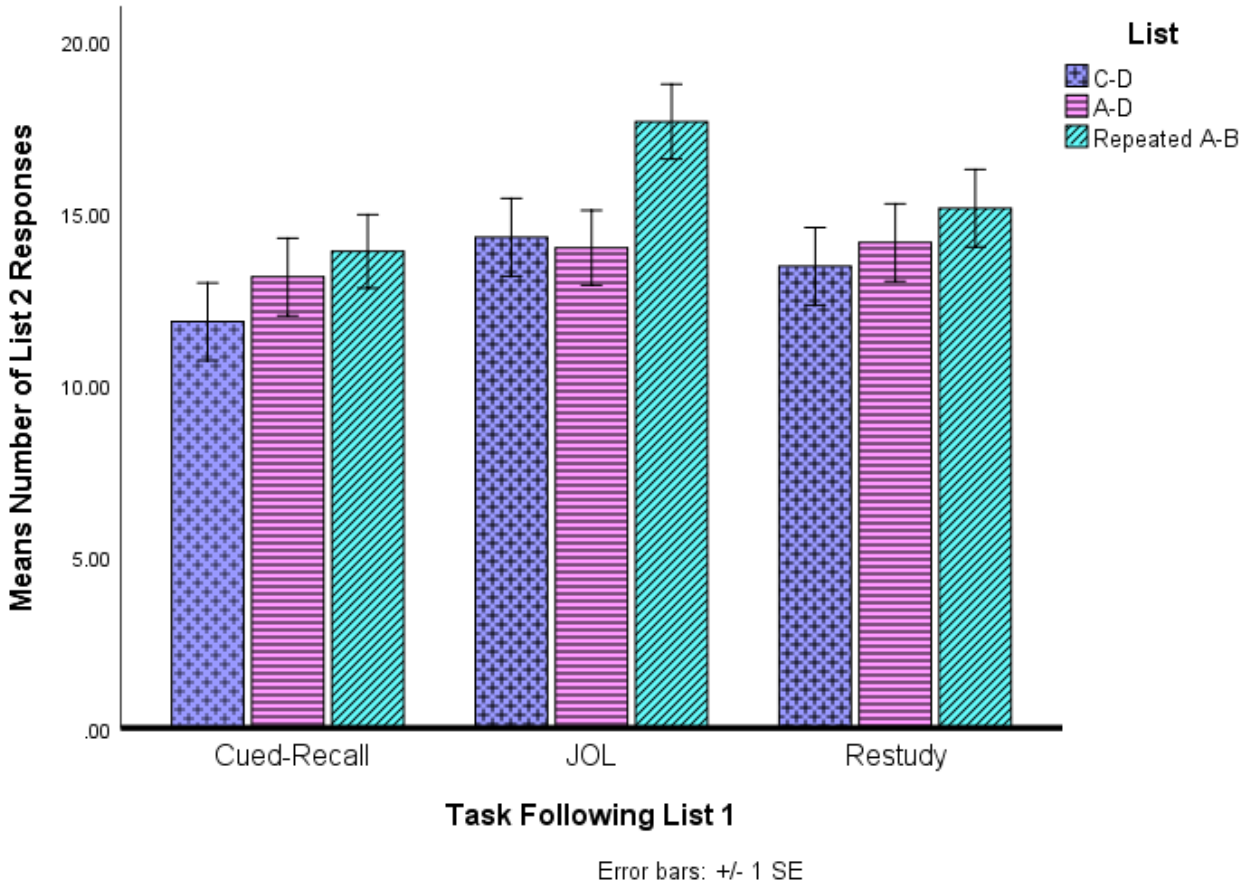

Figure 8. Experiment 1: Mean Number of List 2 Responses (Maximum = 24).

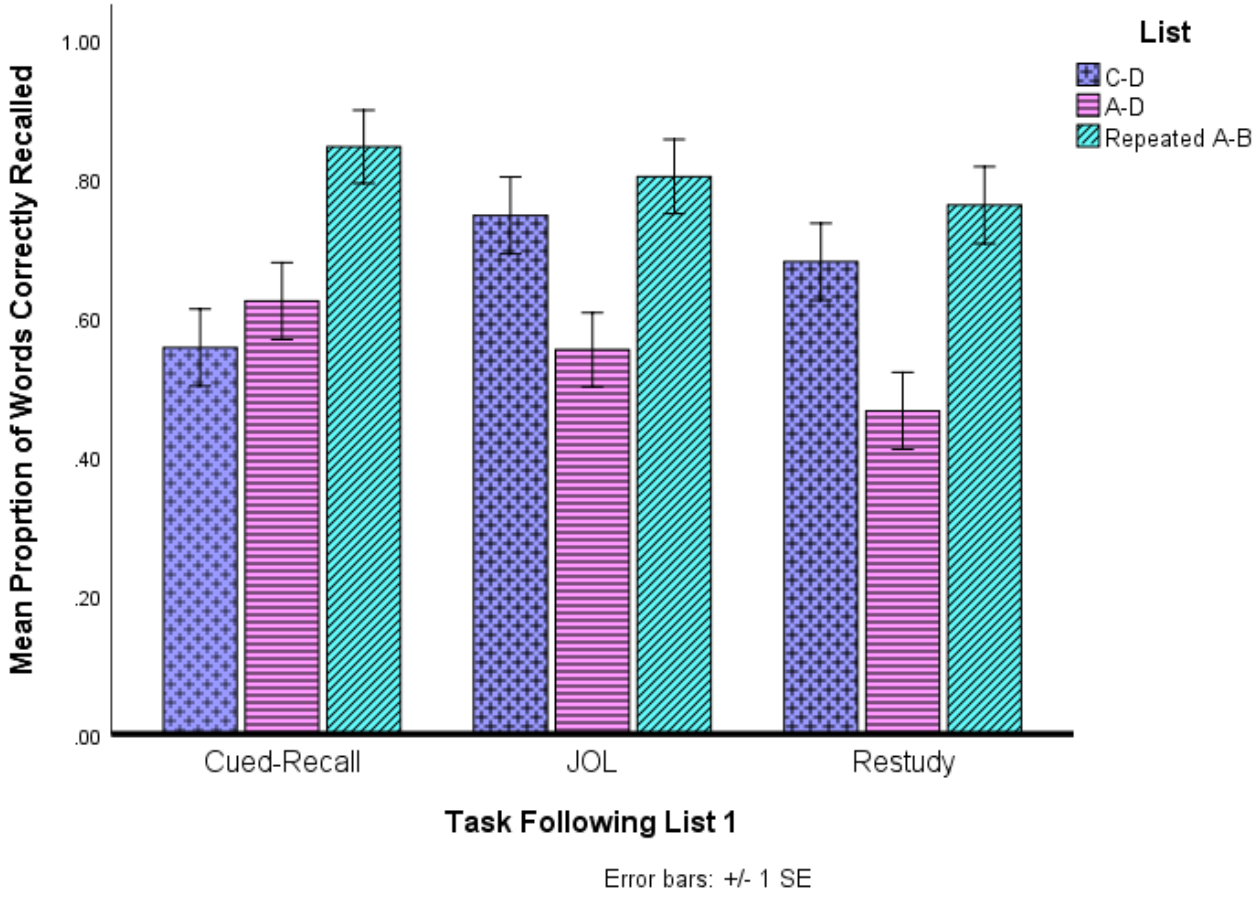

Figure 9. Experiment 1: Mean Proportion of Correct List 2 Responses.

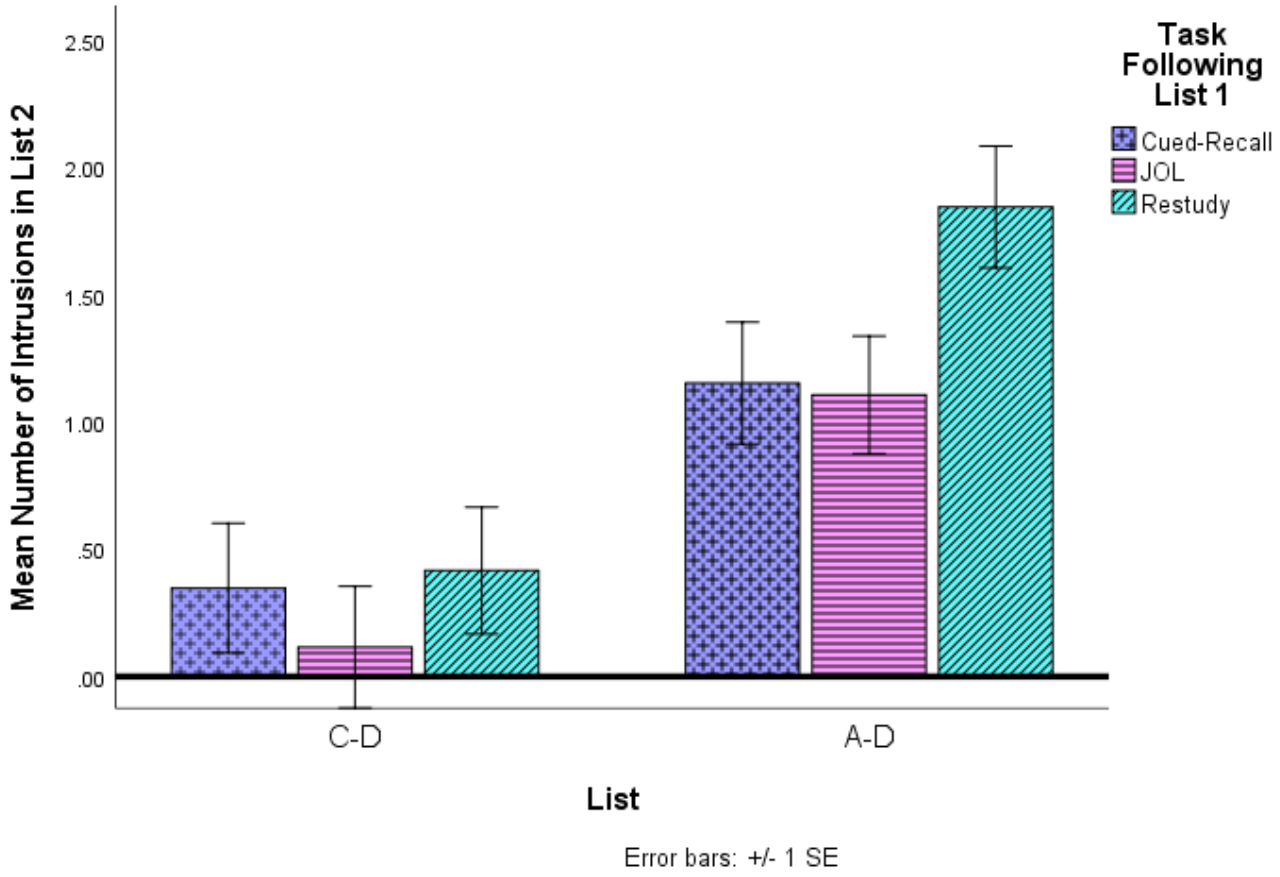

Figure 10. Experiment 1: Mean Number of Intrusions for A-D and C-D items (Maximum = 24).

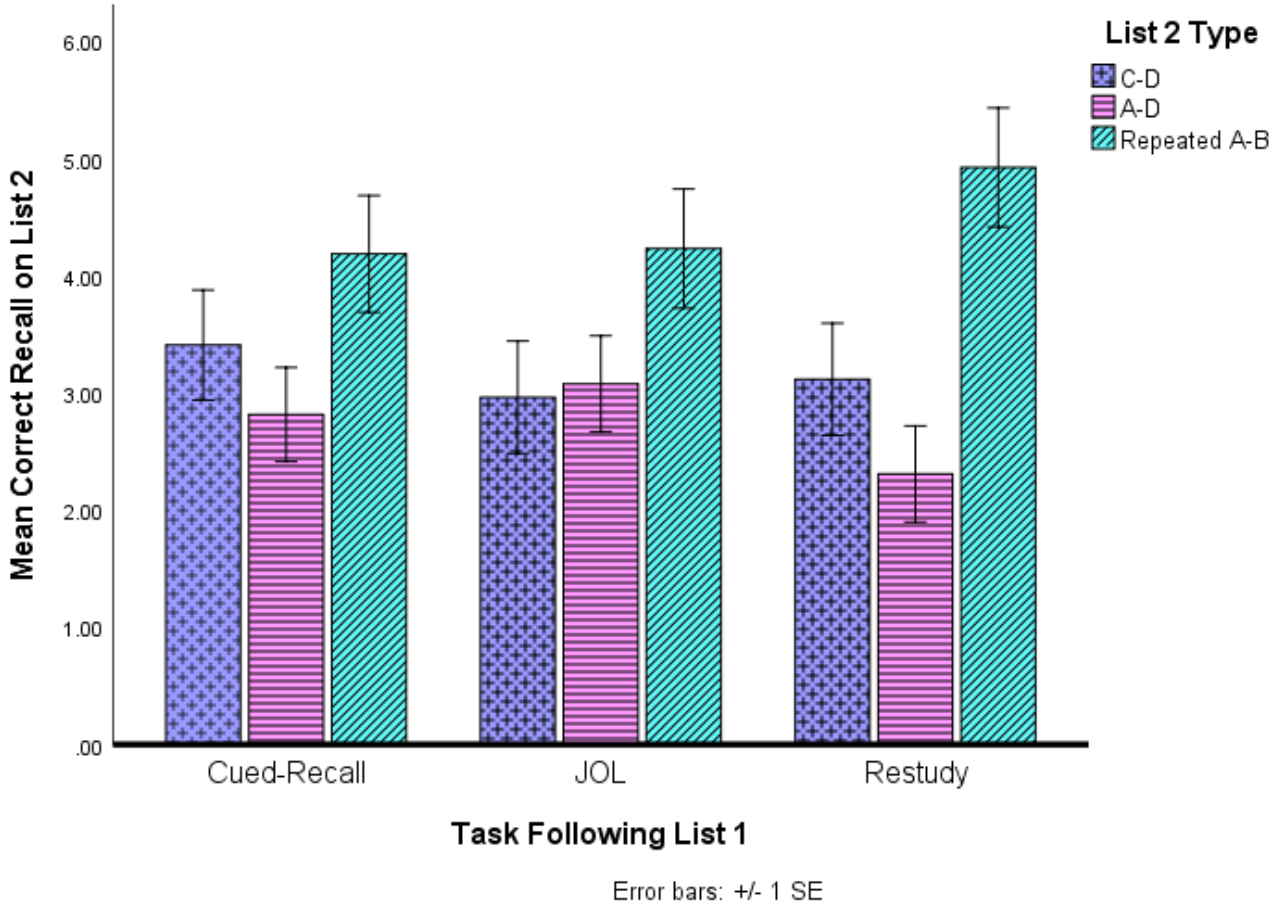

Figure 11. Experiment 2: Mean Number of Correct List 2 Items (Maximum = 10).

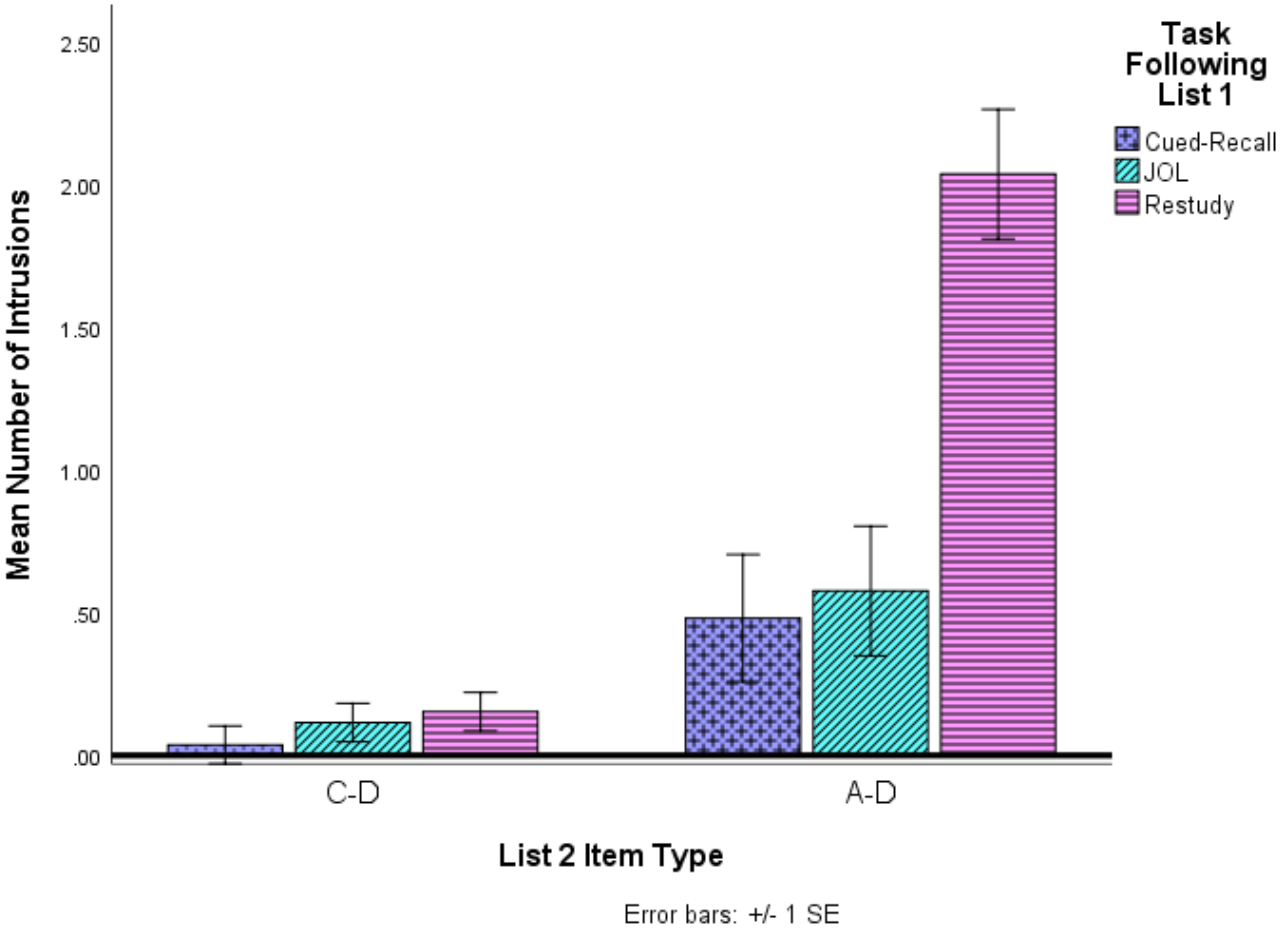

Figure 12. Experiment 2: Mean Number of Intrusions for A-D and C-D items (Maximum = 10).

Appendix A

Experiment 1: Words pairs used in the study.

Version 1			Version 2		
AB	AD	CD	AB	AD	CD
Art-Ranch	Art-Text	Ball-Text	Ranch-Text	Ranch-Lake	Ball-Lake
Lake-Song	Lake-Horn	Chest-Gas	Song-Beard	Song-Barn	Chest-Fruit
Depth-Rice	Depth-Suite	Lane-Card	Rice-Suite	Rice-Ear	Lane-Eat
Ear-Beach	Ear-Pot	Hall-Pot	Beach-Pot	Beach-Depth	Hall-Depth
Sea-Yard	Sea-Fist	Grain-Fist	Yard-Tool	Yard-Fruit	Grain-Fact
Fruit-Pond	Fruit-Tool	Dirt-Tool	Pond-Horn	Pond-Sea	Dirt-Sea
Porch-Bird	Porch-Myth	World-Suite	Bird-Myth	Bird-Hair	World-Hair
Hair-Fort	Hair-Poem	Vein-Poem	Fort-Poem	Fort-Porch	Vein-Porch
Scene-Mouth	Scene-Beard	Tree-Beard	Mouth-Guest	Mouth-Jazz	Tree-Jazz
Jazz-Sky	Jazz-Night	Law-Night	Sky-Verse	Sky-Scene	Law-Scene
Barn-Glass	Barn-Verse	Jet-Verse	Glass-Night	Glass-Meat	Jet-Meat
Meat-Shirt	Meat-Door	Hat-Myth	Shirt-Town	Shirt-Art	Hat-Art
Oil-Teeth	Oil-Feet	Golf-Feet	Teeth-Gas	Teeth-Path	Golf-Path
Path-Disk	Path-Gas	Wood-Joy	Disk-Feet	Disk-Day	Wood-Day
Gift-Hill	Gift-Town	Belt-Town	Hill-Fist	Hill-Oil	Belt-Oil
House-Fig	House-Mud	Coal-Mud	Fig-Mud	Fig-Gift	Coal-Gift
Fact-Corn	Fact-Noise	Desk-Noise	Corn-Noise	Corn-Sheet	Desk-Sheet
Sheet-Tube	Sheet-Wine	Flesh-Door	Tube-Wine	Tube-Fact	Flesh-House
Tea-Roof	Tea-Road	Cell-Horn	Roof-Road	Roof-Van	Cell-Van
Van-Brain	Van-Food	Trial-Road	Brain-Food	Brain-Tea	Trial-Tea
Day-Blood	Day-Card	Faith-Wine	Blood-Card	Blood-Fate	Faith-Cent
Cloth-Goal	Cloth-Crime	Mood-Clay	Goal-Crime	Goal-House	Mood-Barn
Fate-Boat	Fate-Bread	Birth-Bath	Boat-Bread	Boat-Cent	Birth-Fate
Cent-Folk	Cent-Guest	Bench-Hour	Folk-Door	Folk-Cloth	Bench-Cloth

Experiment 2: Word pairs used in the study. A-B/A-D/C-D were combined to form a mixed second list which was split into three groups, one of which would be presented to the participants (10 word-pairs were from each of the A-B, A-D and C-D lists).

Version 1

AB Recall	Mixed Second List 1	Mixed Second List 2	Mixed Second List 3
Leave-Race	AB 1.Give-Arrow	AD 1.Trim-Red	CD 1.Like-Gear
Goal-Lawn	AD 2.Leave-Spicy	AB 2.Leave-Race	AD 2.Berry-Chair
Car-Hedge	CD 3.High-Peach	AD 3.Beach-Food	CD 3.Blow-Feast
Dust-Feel	AD 4.Mix-Fast	CD 4.Blow-Feast	AB 4.Trim-Glide
Law-Dancer	AB 5.Carry-Image	AB 5.Law-Dancer	CD 5.Lock-Envy
Stage-Obey	CD 6.Lock-Envy	CD 6.Hole-Raise	AB 6.Quiet-Handle
Show-Cake	AB 7.Pass-Wax	AD 7.Room-Add	AD 7.Carry-Nice
Mix-Fashion	CD 8.River-Safe	AB 8.Break-Ballet	AB 8.Pain-Island
Break-Ballet	AD 9.Law-Legs	CD 9.High-Peach	AD 9.Music-Think
Soap-Cast	AB 10.Music-Barn	AD 10.Fire-Touch	CD 10.River-Safe
Berry-Mild	CD 11.Blow-Feast	AB 11.Goal-Lawn	AD 11.Door-Sea
Give-Arrow	AD 12.Break-Model	CD 12.Lock-Envy	AB 12.See-Nuts
Music-Barn	AD 13.Dust-Order	AD 13.Quiet-Float	CD 13.High-Peach
Door-Team	CD 14.Time-Moth	AD 14.Pretty-Opera	AD 14.Effort-Ring
Effort-Saw	AB 15.Paper-Gain	AB 15.Dust-Feel	AB 15.Pretty-Peace
Pass-Wax	AD 16.Goal-Movie	CD 16.Pie-Moat	CD 16.Bike-Long
Paper-Gain	CD 17.Hole-Raise	AB 17.Show-Cake	AB 17.Beach-Alarm
More-Pack	AD 18.Soap-Trip	CD 18.Like-Gear	CD 18.Meal-Nose
Smooth-Advice	AB 19.Smooth-Advice	AD 19.Pain-Bee	AB 19.Cut-Dark
Carry-Image	AD 20.Stage-Walk	AB 20.Car-Hedge	AD 20.Smooth-Vision
Trim-Glide	CD 21.Meal-Nose	CD 21.River-Safe	AB 21.Room-Ticket
See-Nuts	AB 22.Door-Team	AB 22.Soap-Cast	CD 22.Pie-Moat
Room-Ticket	CD 23.Bike-Long	AD 23.Head-Play	AD 23.Give-Grass
Pain-Island	AD 24.Car-Seek	CD 24.Time-Moth	AD 24.Pass-Calm
Beach-Alarm	CD 25.Like-Gear	CD 25.Bike-Long	AB 25.Fire-Elegant
Fire-Elegant	AB 26.Effort-Saw	AD 26.Cut-Opinion	AD 26.Paper-Sharp
Head-Quest	CD 27.Pie-Moat	AB 27.Mix-Fashion	CD 27.Hole-Raise
Quiet-Handle	AB 28.More-Pack	CD 28.Meal-Nose	AD 28.More-Scary
Pretty-Peace	AD 29.Show-Blade	AB 29.Stage-Obey	AB 29.Head-Quest
Cut-Dark	AB 30.Berry-Mild	AD 30.See-Sale	CD 30.Time-Moth

Version 2

AB	Mixed Second List 1	Mixed Second List 2	Mixed Second List 3
Socks-Teeth	CD 1.Roast-Flood	CD 1.Drown-Ham	AB 1.Maid-Yogurt
Run-Crisp	AD 2.Socks-Close	CD 2.Roast-Flood	CD 2.Burn-Field
Drop-Invalid	AB 3.Chest-Bait	AB 3.Life-Candy	AD 3.Baked-Clothes
Fruit-Daisy	AD 4.Eyes-Speed	AD 4.Maid-Green	CD 4.Skin-Heart
Eyes-Hero	AD 5.Eat-Snap	CD 5.Hockey-Glow	AB 5.Right-Wood
Brave-Apron	AB 6.Stove-Attack	AD 6.Yellow-Part	CD 6.Fever-Alone
Life-Candy	AD 7.Fruit-Love	AD 7.Good-Plate	AB 7.Lap-Trees
Apple-Haul	AD 8.Bacon-Devil	AB 8.Socks-Teeth	AD 8.Drum-Hurt
Bacon-Intent	CD 9.Hockey-Glow	AD 9.Defend-Hit	AD 9.Stove-Loosen
Eat-Dare	AD 10.Life-Milk	CD 10.Skin-Heart	AB 10.Super-Mouth
Lap-Eggs	AB 11.Drum-Gate	AB 11.Eat-Dare	CD 11.Drown-Ham
Hook-Hairy	CD 12.Skin-Heart	CD 12.Fever-Alone	AD 12.Drag-Man
Drum-Gate	CD 13.Drown-Ham	AB 13.Run-Crisp	AB 13.Smog-Owl
Baked-Bang	AB 14.Brush-Unit	AD 14.Right-Shrimp	AD 14.Swing-Night
Brush-Unit	CD 15.Burn-Field	CD 15.Pump-Item	CD 15.Tip-Hay
Meter-Quick	AD 16.Drop-Store	AB 16.Friut-Daisey	AB 16.Good-Fasten
Stove-Attack	CD 17.Pump-Item	AD 17.Super-Open	CD 17.Afraid-Date
Swing-Beans	AB 18.Lap-Eggs	AD 18.False-Clean	AD 18.Hook-Wrong
Chest-Bait	CD 19.Month-Felt	AB 19.Bacon-Intent	AB 19.False-Foggy
Drag-Napkin	AB 20.Drag-Napkin	CD 20.Tip-Hay	CD 20.Pump-Item
Super-Mouth	CD 21.Afraid-Date	AB 21.Drop-Invalid	CD 21.Hockey-Glow
Good-Fasten	AD 22.Brave-Gloves	AD 22.Smog-Take	AB 22.Yellow-Okay
Yellow-Okay	AB 23.Meter-Quick	AB 23.Eyes-Hero	CD 23.Roast-Flood
Breakfast-Fate	AD 24.Run-Rose	CD 24.Month-Felt	AD 24.Meter-Not
False-Foggy	CD 25.Tip-Hay	CD 25.Afraid-Date	AD 25.Chest-Will
Smog-Owl	AB 26.Hook-Hairy	AD 26.Buckle-Furry	AB 26.Defend-Rain
Maid-Yogurt	AD 27.Apple-Shower	AB 27.Brave-Apron	AD 27.Brush-Fry
Defend-Rain	CD 28.Fever-Alone	AD 28.Breakfast-Forest	AB 28.Buckle-Pair
Right-Wood	AB 29.Baked-Bang	CD 29.Burn-Field	AD 29.Month-Felt
Buckle-Pair	AB 30.Swing-Beans	AB 30.Apple-Haul	AB 30.Breakfast-Fate

Appendix B

Studying to remember consent form

IMPROVE LIFE.

CONSENT TO PARTICIPATE IN RESEARCH

Research Project Title: Studying to Remember REB# 17-08-035

Purpose

The purpose of this letter is to provide you with the information you require to make an informed decision on participating in this research. The goal of the research project is to understand the influence of methods of studying on the ability to remember verbal information. The researcher will undertake to publish the results of the project in a peer-reviewed journal. You may request a copy of the averaged results by emailing the principal investigator at the end of the Winter 2018 semester. No individual is identified in the stored data as the goal of the study is to compare averaged results across experimental conditions.

Principal Researcher:

Dr. Harvey Marmurek, Professor, Department of Psychology
(hmarmure@uoguelph.ca**), x53673**

Eligibility:

Any student registered in a University of Guelph Psychology course offering research participation credits may participate in this study for 0.5 credit.

Procedures:

If you volunteer to participate in this study, you will:

1. Study a 20-item list of common words followed by a picture projected on a screen.

2. Write down the words on the list or draw the picture.

3. Study a second list of common words.

4 Recall the second list of items.

Each participant will receive only one of the tasks at step 2 so that we may determine the influence of studying and testing on memory for the second list of words.

The total length of time for participation is about 20 minutes.

Benefits

While there is no direct benefit to you, your participation is important because the results will increase the knowledge of how testing improves learning over and above mere studying.

Risks and Confidentiality

There are no known risks associated with this study.

No identifying information will be stored with the memory data collected in the study. The memory data you provide will be stored anonymously on laboratory computers that are accessible to only qualified laboratory personnel under the supervision of Dr. Marmurek. The hard copies of your written responses will be shredded once all data have been entered on the computer. The electronic data you provide will be kept for 10 years, and may be used in future studies to answer similar research questions.

Withdrawal from the Study

Participation in this study is voluntary. You may refuse to participate, refuse to answer any questions or withdraw from the study at any time with no effect on your future academic status. If you decide to withdraw from the study, you can request to have your data withdrawn. In the event you decide to withdraw from the study, any data you have provided will be shredded.

Ethics Approval:

Please feel free to contact Dr. Marmurek (hmarmure@uoguelph.ca) with any questions you might have about the research study.

This project has been reviewed by the Research Ethics Board for compliance with federal guidelines for research involving human participants

*If you have questions regarding your rights and welfare as a research participant in this study (REB#*17-08-035*), please contact: Director, Research Ethics; University of Guelph; reb@uoguelph.ca; (519) 824-4120 (ext. 56606).*

Name of Participant (please print) _____

Signature of Participant: _____

Date: _____

Studying to remember debriefing form

DEBRIEFING FORM

IMPROVE LIFE.

Project Title: Studying to Remember

Faculty Researcher: Dr. Harvey Marmurek, Department of Psychology

PURPOSE OF THE STUDY

The purpose of the research is to investigate the role of studying and testing on memory for word lists.

In this study you were asked to learn two lists of common words. Each list was followed by a picture. Following the first list, you were asked either to recall the words in the list or to draw the picture. The task (word recall or picture drawing) was varied randomly across participants.

The main hypotheses for this study are as follows: (1) testing the first list will facilitate learning and memory of the second list relative to groups that drew the picture after studying the first list. The results will have implications for optimizing the way testing procedures should be used to improve learning.

Note that the purpose of the study is **NOT** to assess an individual's memory capacity in general. Rather, we are interested in how the manipulations of study and testing condition influence learning and memory.

CONTACT INFORMATION FOR RESULTS

Participants may contact the faculty researcher, Dr. Harvey Marmurek, by phone (519-824-4120 ext. 53673) or email (hmarmure@uoguelph.ca) if they wish to receive information summarizing the results of the study at the end of the semester. Only group averages for each experimental condition will be provided. No individual results are given out as researchers do not know the name of the participant who provided the data.

Appendix C

Materials presented to participants

Word recall task example:

1. Please write the word that accompanied each of the following words from the list that you studied. You have 2.5 minutes to complete this task.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____

Judgement of learning task example:

1. Using a scale of 0-100, with 0 being completely uncertain and 100 being completely certain, please indicate how strongly you feel that you can accurately recall each of the words that are missing in the pairings that you studied. You have 2.5 minutes to complete this task.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____

Picture drawing task example (only ever shown as a potential task to be completed):

1. Please redraw the picture you were shown from memory. You have 2.5 minutes to complete this task.

Appendix D

Metamemory and learning consent form

IMPROVE LIFE.

CONSENT TO PARTICIPATE IN RESEARCH

Research Project Title: Metamemory and Learning REB# 17-09-004

Purpose

The purpose of this letter is to provide you with the information you require to make an informed decision on participating in this research. The goal of the research project is to understand whether judging how well one knows recently learned information predicts learning of new knowledge. The researcher will undertake to publish the results of the project in a peer-reviewed journal.

Principal Researcher:

Dr. Harvey Marmurek, Professor, Department of Psychology

(hmarmure@uoguelph.ca**), x53673**

Eligibility:

Any student registered in a University of Guelph Psychology course offering research participation credits may participate in this study for 0.5 credit.

Procedures:

If you volunteer to participate in this study, you will:

1. Study a list of common words projected on a screen.

2. Just how well you know the words or draw a picture.

3. Study a second list of common words.

4. Recall the second list of items.

Each participant will receive only one of the tasks at step 2 so that we may determine the influence of knowledge judgments on memory for the second list of words.

The total length of time for participation is about 25 minutes.

Benefits

While there is no direct benefit to you, your participation is important because the results will increase the knowledge of how testing improves learning over and above mere studying.

Risks and Confidentiality

There are no known risks associated with this study.

No identifying information will be stored with the memory data collected in the study. The memory data you provide will be stored anonymously on laboratory computers that are accessible to only qualified laboratory personnel under the supervision of Dr. Marmurek. The hard copies of your written responses will be shredded once all data have been entered on the computer. The electronic data you provide will be kept for 10 years, and may be used in future studies to answer similar research questions.

Withdrawal from the Study

Participation in this study is voluntary. You may refuse to participate, refuse to answer any questions or withdraw from the study at any time with no effect on your future academic status. If you decide to withdraw from the study, you can request to have your data withdrawn.

Ethics Approval

Please feel free to contact Dr. Marmurek (hmarmure@uoguelph.ca) with any questions you might have about the research study.

This project has been reviewed by the Research Ethics Board for compliance with federal guidelines for research involving human participants

*If you have questions regarding your rights and welfare as a research participant in this study (REB#*17-09-004*), please contact: Director, Research Ethics; University of Guelph; reb@uoguelph.ca; (519) 824-4120 (ext. 56606).*

Name of Participant (please print) _____

Signature of Participant: _____

Date: _____

Metamemory and learning debriefing form

DEBRIEFING FORM

Project Title: Metamemory and Learning

Faculty Researcher: Dr. Harvey Marmurek, Department of Psychology

PURPOSE OF THE STUDY

The purpose of the research is to investigate the role of studying and retrieval on memory for word lists.

In this study you were asked to learn two lists of common words. Each list was followed by a picture. Following the first list, you were asked either to judge how well you would remember the words in the list or to draw the picture. The task (judgment of memory or picture drawing) was varied randomly across participants.

The main hypotheses for this study are as follows: (1) judgments about the first list will facilitate learning and memory of the second list relative to groups that drew the picture after studying the first list. The results will have implications for optimizing the way study procedures should be used to improve learning.

Note that the purpose of the study is **NOT** to assess an individual's memory capacity in general. Rather, we are interested in how the manipulations of study conditions influence learning and memory.

CONTACT INFORMATION FOR RESULTS

Participants may contact the faculty researcher, Dr. Harvey Marmurek, by phone (519-824-4120 ext. 53673) or email (hmarmure@uoguelph.ca) if they wish to receive information summarizing the results of the study at the end of the semester. Only group averages for each experimental condition will be provided. No individual results are given out as researchers do not know the name of the participant who provided the data.