**Disruption of the Relational and Item-specific Processing Supports the Negative Outcomes of Multiple-Choice Testing with Additional Lures**

by

Bavani Paneerselvam

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2017

Keywords: testing effect, retrieval, types of processing, multiple-choice test



Pending Approval by
Aimee Callender, Associate Professor of Psychology
Alejandro Lazarte, Associate Professor of Psychology
Ana Franco-Watkins, Associate Professor of Psychology
Dan Svyantek, Associate Professor of Psychology

Abstract

Multiple-choice retrieval practice with additional lures reduces retention on a later test (Roediger & Marsh, 2005). However, the mechanism underlying the negative outcomes with additional lures is poorly understood. Given that the positive outcomes of retrieval practice are associated with enhanced relational and item-specific processing (Zaromb & Roediger, 2010), it is plausible that the negative outcomes are due to disrupted relational and item-specific processing. The main goal of this dissertation was to examine whether the negative outcomes are related to disruption of the relational and item-specific processing. Experiment 1 compared the relational and item-specific processing in a non-inclusive intervening multiple-choice test (1 lure versus 5 lures), whereas experiment 2 compared these processing in an inclusive multiple-choice test (correct none of the above versus wrong none of the above). Across two experiments, results show that retrieval practice with additional lures disrupts both the relational and item specific processing resulting in reduced retention.

Table of Contents

List of Tables

Appendix

List of Figures

Experiment 1



Experiment 2

## Introduction

For more than 100 years now, cognitive psychologists have been developing and evaluating the efficacy of learning techniques due to the important implications of these techniques for classroom learning and self-regulated learning. Among the techniques that have gained considerable attention in the laboratory is retrieval practice (i.e., *the testing effect*). Due to the numerous and reliable memorial benefits found in testing in laboratories, this learning technique has been recommended as an instructional technique in educational settings for similar advantages. Although testing improves memory outcomes, the memorial benefits associated with multiple-choice testing has been inconsistent. As a result, the existent theories of the testing effect are insufficient to provide a unifying explanation for the various outcomes associated with the multiple-choice testing.

Recent studies show that the relational processing and item-specific processing hold possibilities in explaining the various outcomes of testing. But no studies have examined the influence of type of processing on testing with multiple-choice questions. The aim of this dissertation is to examine the cognitive mechanisms underlying the negative outcomes of multiple-choice testing with additional lures using the relational and item-specific theoretical framework.

### Testing Effect

The testing effect paradigm compares a Study-Study-Test condition against a Study-Test-Test condition where the repeated testing condition outperforms the repeated studying condition on the final test. The benefit of repeated testing is termed as the *testing effect* (positive testing effect) which is typically observed when the final test is administered after a delay (Roediger &

Karpicke, 2006a) and can be enhanced with feedback (e.g., Butler, Kapicke, Roediger, 2007; Kang, McDermott, & Roediger, 2007).

The robustness of this phenomenon has been reported with a wide variety of stimuli (e.g., Hogan & Kintsch, 1971; Wheeler & Roediger, 1992), across domains (e.g., Carpenter, Pashler & Cepeda , 2009) and ages (e.g., Meyer & Logan, 2013). Furthermore, testing facilitates retrieval of non-tested questions (e.g., Chan, McDermott, & Roediger, 2006) and potentiates further studying (e.g., Arnold & McDermott, 2013). These benefits of the testing effect also generalize to various test formats, both in the laboratory and the classroom (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011).

Although repeated retrieval enhances retention, some studies show that the benefit depends largely on the test formats. For example, initial testing with short-answer benefits retention compared to multiple-choice (e.g., Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; however see Smith & Karpicke, 2014). Subsequent studies, however, reveal that the multiple-choice test format increases retention more than the other test formats (e.g., Little, Bjork, Bjork, & Angello, 2012; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). Furthermore, some argue that the benefit of testing is not dependent on the test format per se, but on the successful retrieval of the correct answer on the intervening test (Smith & Karpicke, 2014). Although some multiple-choice questions can induce successful retrieval of the correct answer on the intervening test resulting in the positive testing effect (e.g., Little & Bjork, 2010), others can reduce the positive testing effect (e.g., Roediger & Marsh, 2005). These divergent outcomes are due to the manipulation of number of lures on the intervening test.

**Testing Effect with Multiple-Choice Test**

Although studies on multiple-choice testing were already known in the early 20th century (Spitzer, 1939), only recently multiple-choice testing has gained considerable attention (Little & Bjork, 2010, 2011). In a typical multiple-choice testing effect paradigm, participants read a set of prose passages and take an intervening multiple-choice test that consists of questions from both the passages that were read and also from passages that were not read (control). The final test can be of the same format (i.e., multiple-choice) as the intervening test or in a cross-format (i.e., free-recall and cued-recall) and they both contain the previously tested questions and non-tested control questions. The general finding shows that taking an intervening multiple-choice test can benefit retention compared to the non-tested control questions, in both the same-format and cross-format final tests, resulting in the *positive testing effect* (e.g., Kang, McDermott, & Roediger, 2007).

Although intervening multiple-choice testing can aid in retention, it can also have negative side effects. One problem of the multiple-choice test is that it exposes students to misinformation in the form of lures when they are reading through the correct and incorrect (i.e., lures) answer options (e.g., Roediger & Marsh, 2005). Believing the misinformation to be true is termed as the *negative suggestion effect* (Remmers & Remmers, 1926) and is likely to occur due to repeated exposure to the misleading statements (Hasher, Goldstein, & Toppino, 1977). The negative suggestive effect has been demonstrated by judgment ratings. For example, participants rated incorrect lures that reappeared from the initial true-false test and multiple-choice test as more truthful than novel lures (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). The negative suggestion effect also extends to answer production where participants use the incorrect answer options (i.e., incorrect lures) from the intervening multiple-choice test as answers on final

multiple-choice and cued-recall tests (e.g., Odegard & Koen, 2007). The increased intrusion of lures from the previously tested questions compared to the non-tested control questions is termed as the *negative side effect of testing*

Critically, the intrusion of lures on the final test is shown to emerge due to faulty reasoning rather than exposure (e.g., Huelser & Marsh, 2006; Marsh, Roediger, Bjork, & Bjork, 2007). If it was just the exposure to the incorrect lures then the incorrect lures should intrude on the final tests regardless of whether the students choose the correct answer or the incorrect lures on the intervening test. However, studies show that the intrusion of lures occurs only when the students commit to the incorrect lures on the intervening test (e.g., Roediger & Marsh, 2005). Furthermore, even if students choose *lure A* on the intervening test, they do not change it to *lure B* on the final test (Butler, Marsh, Goode, & Roediger, 2006; Roediger & Marsh, 2005). Therefore, committing to the lures through the reasoning process increases intrusions.

Taken together, multiple-choice testing can have both positive and negative outcomes and the different consequences on the final test are largely due to different levels of performance on the intervening multiple-choice testing. Further, the types of answer options on the intervening test can either enhance or attenuate the positive and negative testing effects.

**Multiple-Choice Test with different Types of Answer Options**

The testing effect literature on multiple-choice testing with the standard number of alternatives (i.e., 4 alternatives; 1 target and 3 lures) has consistently shown both the positive and negative side effect of testing, critically with the positive testing effect outweighing the negative side effect (e.g., Marsh, Bjork & Bjork, 2006; Marsh & Roediger, 2006)

Studies that used non-standard answer options (e.g., additional lures) have yielded similar results (e.g., Roediger & Marsh, 2005). For example, Roediger and Marsh (2005) examined how prior testing, with additional lures (i.e., 2 lures, 4 lures, 6 lures) and without feedback, affected

the immediate final test performance. Their results indicated that the additional lures decreased the intervening multiple-choice test accuracy, however, increased selection of the correct answers on the final test, thereby resulting in an overall positive testing effect. However, the lure intrusion increased on the final test with the increase of lures on the intervening test, resulting in the negative testing effect. That is, for the previously tested questions, compared to the control questions (questions that were not initially tested), participants produced more lures as the answers on the final test (these questions had additional lures) (see also, Butler et al., 2006). In sum, taking a multiple-choice test with additional lures can be both beneficial and harmful, with the benefit outweighing the cost.

Although in these answer options students select the correct answer from a number of incorrect options (lures), instructors sometimes use another type of non-standard answer option where either all or none of the answer options could be the correct answer. Such inclusive answer options can be of two types namely the *none-of-the-above* option and the *all-of-the-above* option.

Odegard and Koen (2007) extended upon Roediger and Marsh's study and examined the testing effect of the *none-of-the-above* (NOTA) answer option on the intervening multiple-choice test without feedback. The intervening multiple-choice question had two types of answer options: the standard format (4 options) and the inclusive format (NOTA as one of the five options). The inclusive format was further broken down into the correct-NOTA (NOTA is the correct answer) and the incorrect-NOTA (NOTA is not the correct answer). Thus, the correct-NOTA exposes students only to the lures, whereas the incorrect-NOTA and the standard format expose them both to the lures and the target. They found that prior testing with only lures, which in this case is the correct-NOTA, attenuated the positive testing effect on the final cued-recall

(Experiment 1) and the 4-option standard format MCQ tests (Experiment 2). That is, the final test recall for the initially tested correct-NOTA that had only the lures was significantly lower than the recall on the control questions. Also, the lure intrusion was the highest for the correct-NOTA, resulting in the negative side effects. Critically, the finding that prior testing with only the lures in the correct-NOTA negated the positive testing effect is concerning because generally the positive testing effect is shown to outweigh the negative side effects (Marsh, Roediger, Bjork, Bjork, 2007). Therefore, the correct-NOTA questions can be harmful because they expose students to only the misinformation that can be incorporated into their memory (see also Jang, Pashler & Huber, 2014).

If the presence of lures (i.e., correct-NOTA) on an intervening multiple-choice test negates the testing effect then the absence of lures in the all-of-the-above question that has only the correct options (correct-AOTA) should reverse the outcome, resulting in the positive testing effect. Bishara and Lanzo (2014) examined this hypothesis on a final cued-recall test (Experiment 1) and a final inclusive format MCQ test with correct-AOTA and wrong-AOTA options (Experiment 2). Participants took an intervening test that had questions in the standard format (3 incorrect lures and a correct answer) and inclusive options; correct-AOTA (no lures), and wrong-AOTA (4 incorrect lures and a correct answer). The final cued-recall test and the multiple-choice test in the inclusive format had the same question stems from the intervening test. Results show that prior testing with the correct-AOTA which had no lures enhanced the testing effect and reduced lure intrusions. Similar benefit has also been found for the final MCQ test in a standard format (4 options; 3 incorrect lures and 1 correct answer) (Paneerselvam & Callender, 2016).

To summarize, incorrect lures on the intervening test can have different outcomes on the final test: a positive testing effect (Bishara & Lanzo, 2015), a negative testing effect (Odegard & Koen, 2007), and negative side effects of testing (i.e., lure intrusion) (Roediger & Marsh, 2005). With regard to the inclusive multiple-choice tests, the presence of only the incorrect lures (correct-NOTA) on the intervening test negates the positive testing effect and increases the negative side effects whereas the absence of incorrect lures (correct-AOTA) reverses this pattern. Additionally, the positive testing effect was observed even when no feedback was given following the intervening test.

**Theories of the Testing Effect**

One theory suggests that the positive testing effect arises from retrieval effort involved in the practice testing (e.g., Bjork, 1975; Bjork & Bjork, 1992; Jacoby, 1978; Karpicke & Roediger, 2010; Pyc & Rawson, 2009). According to the *retrieval effort theory* (Bjork, 1994), difficult intervening tests demand more effort to retrieve the correct answer, thus the difficult but successful retrieval benefits retention of the tested materials. Evidence for this theory is provided by retrieval cue studies in which participants' retention was higher on the final test for the questions on the intervening test that had fewer cue (e.g., Carpenter & DeLosh, 2006). Also, increasing the retention interval between the study and intervening test, thereby increasing the difficulty of the test, has shown to result in the positive testing effect (e.g., Pyc & Rawson, 2009). According to this theory, the additional lures on the intervening multiple-choice test should increase difficulty of the questions and improve retention on the final test. However, Roediger and Marsh (2005) study failed to support this conclusion; additional lures decreased the positive testing effect and increased the negative side effect. Similarly, the correct *none-of-the-above* questions with only lures should be difficult and require more effort to retrieve the correct

answer. However, Odegard and Koen (2007) found the negative testing effect for the *none-of-the-above* questions which is presumably more difficult than questions that had a single correct answer.

According to the *transfer appropriate processing theory* (TAP), both the practice test and final test involve similar processing, whereas the processing differs across restudying and final tests (Morris, Bransford & Franks, 1977). Thus, the overlap in the processing in the testing condition improves memory (Roediger & Karpicke, 2006b). For example, Johnson and Mayer (2009) showed participants a multimedia presentation about lightning. Later participants restudied presentation or took an intervening test. The intervening test was further divided into retention test ("*write down an explanation of how lightning works*") or a transfer test ("*suppose you see clouds in the sky but no lightning, why not?*"). After a delay of 1 week, participants took a final retention test or a final transfer test. Consistent with the TAP theory, they found higher retention when the intervening test type and the final test type matched. Similar results have been found for match in the test formats (e.g., Nungster & Duchastel, 1982) and for identical questions (as opposed to rephrased questions) (e.g., McDaniel and Fisher, 1991). According to this theory, the positive testing effect should be enhanced for the intervening multiple-choice testing because the processing at intervening test aligns with the processing on the final test, as opposed to the misaligned processing between the restudy and the final test. Counter to this prediction, prior testing with multiple-choice with additional lures attenuates the positive testing effect in Roediger and Marsh (2005) study. Similarly for the inclusive NOTA questions, although the processing in the intervening test overlapped with the processing in the final test, retention was lower for the testing condition compared to the control condition (non-overlapping processing). Even when the intervening test and the final test was a multiple-choice test

(Experiment 2), which presumably has more overlaps in the processing, the control condition had higher retention than the testing condition.

The *interference theory* posits that the increase in the number of competing associations would clutter memory of these similar associations, reducing access to the target (e.g., Postman & Underwood, 1973). According to this theory, prior testing with additional lures should lower retention on the final test when previously tested with additional lures. Consistent with this prediction, Roediger and Marsh (2005) showed that increasing the number of multiple-choice decreased the positive testing effect and increased the negative side effect (i.e., lure intrusion). However, other studies showed the positive testing effect (e.g., Whitten & Leonard, 1980) or no effect (e.g., Brown, 1988) when tested with additional lures on the intervening test, failing to support this theory.

The *retrieval blocking* theory posits that retrieval of the misinformation blocks access to the target information (Raaijmakers & Shiffrin, 1981). According to this theory, retrieval of lures on the intervening test should block retrieval of the correct target, resulting in lower retention. Critically, this theory also posits that although the target information is blocked, it is still available and can be resurfaced with cues (e.g., Rundus, 1973). Although there should be lower retention in a non-cued test such as free-recall, retention should improve in a cued test such as multiple-choice. Thus, Odegard and Koen (2007) proposed that the only incorrect lures in the correct *none-of-the-above* questions (all options are incorrect) blocked the retrieval of the correct answer on the final test, resulting in the negative testing effect. However, against prediction, the negative testing effect was robust even in test that provides cues such as the multiple-choice test (Experiment 2).

Taken together, although each of the multiple-choice question types is supported by various theories, the existent theories are limited in two ways. First, the multiple theories of the multiple-choice testing provide a variety of competing views with broad explanations, thus making it difficult to pinpoint the precise underlying mechanism. Second, the theories are limited in explaining either the positive outcome or the negative outcome, failing to provide a unifying account for both the outcomes.

One theory that could explain both the positive and negative effects of testing, providing a unifying theory of the testing effect, is the theory based on the distinction between item-specific and relational processing (Hunt & McDaniel, 1993; Mulligan & Lozito, 2004; Peterson & Mulligan, 2013).

### *Relational and Item-Specific Processing Theories*

Cognitive psychologists have long established the importance of mental organization and distinctiveness on memory (e.g., Köhler, 1941; Mandler, 1967, 1972). An organizational approach traces back its root to chunking, where people remember better when a large piece of information is broken down into chunks (Miller, 1956), more so when the chunks are organized according to categories (Mulligan, 2005). Organization refers to the relationship among the items on a study list (how the items fit together). Deese (1959) found participants recalled 49% of the words from a categorized list compared to 37% from a random list, demonstrating the benefits of organization. Furthermore, just the act of organizing information without an attempt to memorize can improve recall. For example, Mandler (1967) found that the recall on a surprise test was comparable between participants who had previously sorted the words into categories and participants who memorized the words. On the other hand, studies show that distinctiveness helps memory (e.g., Hunt & Mitchell, 1982; McDaniel & Einstein, 1986). Distinctiveness refers

to the differences or uniqueness among the items on a study list, which facilitates a discriminate process (Hunt & McDaniel, 1993). For example, a study list that consist of the items "sheep, *watermelon*, pig, and duck", "*watermelon*" that appears distinctive will be more accurately recognized in a recognition test.

Over the years, various theories have emerged that are linked to organization and distinctiveness, such as the *distinctiveness hypothesis* (Hunt & Einstein, 1981; Hunt & Elliott, 1980), the *organization and distinctiveness view* (Hunt & McDaniel, 1993), *distinctive processing* (Hunt & Worthen, 2006), *multifactor theory* (Mulligan & Lozito, 2004) and more recently, *item/relational processing theory* (Schmidt, 2008). The common property of these theoretical frameworks is the focus on the distinction between item-specific and relational processing (e.g., Hunt & McDaniel, 1993). Relational processing emphasizes organization whereas item-specific processing emphasizes distinctiveness.

*Relational processing* refers to processing of common features among items in a list. For example, making connections between the different words "*cucumber*" and "*tomato*" and knowing that they belong to the category "*vegetable*" makes it easier to recall the words according to that category. In contrast, *item-specific processing* refers to processing of the distinctive features of an individual item. For example, noticing "*cucumber*" is long and "*tomato*" is round enhances the unique features of each item that the word refers to, and as a result the words are easier to differentiate. Therefore, the theoretical framework based on the distinction between item-specific and relational processing contrast the processing of single items (item-specific) with multiple items (relational processing).

Although relational processing and item-specific processing both enhance memory, they work in an opposite direction. This contradiction echoes Hunt and McDaniel's (1993) question

on "*how can both similarity and difference be beneficial to memory?*" A plausible explanation is that relational processing facilitates a retrieval plan (*vegetables were studied*), and item-specific processing helps to identify the specific vegetables (*cucumber and tomato*). Thus, the theoretical framework based on the distinction between the relational processing and item-specific processing posits that memory and learning are improved when both types of processing work in combination (e.g., Einstein & Hunt, 1980; Hunt, 2012; Hunt & McDaniel, 1993). However, some learning techniques only benefit either relational or item-specific processing (e.g., Grimaldi, Poston, & Karpicke, 2014; Lipowski, Pyc, Dunlosky, & Rawson, 2014). Grimaldi et al. (2014) examined if concept mapping enhances item-specific and/or relational processing. In order to capture the processing promoted by the concept map, they compared concept map (unknown processing) with word sorting (relational processing) and pleasantness rating (item-specific processing). They found that concept mapping promoted relational processing (as indicated by high clustering equivalent to the sorting condition) but disrupted item-specific processing (lowest outcome in the recognition test). As a result, these techniques do not always benefit memory under all conditions.

Furthermore, item-specific processing and relational processing affect memory and learning differently depending on test formats (Hunt & Einstein, 1981). Free-recall test relies both on the relational and the item-specific processing, heavily on the former (e.g., Hunt & McDaniel, 1993). For example, relational information is used to identify the interrelations among the words "*cucumber, carrot,* and *tomato"* as members of the category "*vegetable"* (cue), and the self-generated cue helps to narrow down the search set. On the other hand, item-specific information is used to process the unique features of each word in the search set, thereby facilitating the accurate retrieval of the studied words by filtering the competing word "*cherry*".

In contrast, in a cued-recall test where the cues are provided, relational processing is not beneficial. Instead, item-specific information is more useful to retrieve the correct word "*cucumber*" for a given cue "*vegetable*" based on the prior processing of the unique features (e.g., Peterson & Mulligan, 2012). Similarly, in a recognition test where one has to discriminate between the studied word and a non-studied word, item-specific information is useful (Hunt, 2003). This is because the unique features of the words facilitate discrimination of the studied word "*cucumber*" from a non-studied word "*cherry*". Therefore, relational processing operates in free-recall test, whereas item-specific processing operates in cued-recall and recognition tests.

In sum, memory is enhanced when the learning techniques promote both the item-specific and relational processing. However, not all learning techniques promote both types of processing, and different test formats rely differently on the processing.

***Testing effect: Retrieval Processing and Item-specific Processing***

Questions about what accounts for the positive testing effect have been examined using the types of processing associated with retrieval (Congleton & Rajaram, 2012; Knouse, Rawson, Vaughn, & Dunlosky, 2016).

Zaromb and Roediger (2010; experiment 2) examined why the positive testing effect occurs with a free-recall practice test. That is, they examined if a free-recall practice test promoted relational and/or item-specific processing, which would in turn facilitate the final test outcome. In their study, participants studied a categorized words list (words from various taxonomic categories) and completed either a practice free-recall test or restudied (control). Testing, compared to restudying, improved their performance on the final free-recall test and cued-recall. Zaromb and Roediger (2010) associated the positive testing effect with organization (as measured by category clustering, ARC) and distinctiveness. Practice testing (free-recall)

enabled the participants to mentally organize the words into categories and they used these categorical cues to guide retrieval on the final test (free-recall). This is supported by the measurement of relational processing (words from the same category are clustered together during recall). Furthermore, the practice testing promoted item-specific processing (average number of words correctly recalled within categories). Therefore, the final free-recall test that is reliant on the relational processing and the cued-recall test that is reliant on item-specific processing benefitted from the free-recall practice test that promoted both these processing.

Rawson and colleagues (2015) used the *general multifactor account of testing effects*, which is based on the distinction among the item-specific processing, relational processing, and intraitem processing to explain the positive testing effects with word pairs in a final free-recall test. Practice testing with categorized word pairs promotes the cue-target (e.g., *moon-spoon*) intraitem processing, target (e.g., *spoon*) item-specific processing, and category: target-target (e.g., *kitchen: spoon and knife*) interitem relational processing. Because practice testing promotes all types of processing in all test formats, cued-recall practice testing benefits the final free-recall test that is reliant on similar processing (however, see Peterson & Mulligan, 2013). Further support comes from the higher categorical clustering in the testing condition compared to the study condition. Since practice testing promotes all types of processing, the positive testing effect was also observed when both the practice test and final test was in a cued-recall format. Thus, the final cued-recall test that is reliant on the intraitem processing matched the similar processing promoted during the practice testing with a cued-recall test.

Using the *distinctiveness theory*, Wissman and Rawson (2015) have also provided evidence for the enhancement of relational and item-specific processing for retrieval practice that happens collaboratively. In their study, participants studied categorized word list, and took a

recall practice test either collaboratively or individually. Following that, participants individually completed final free-recall test and recognition test. In a series of experiments, they found that collaborative group recalled more than individual participants. Importantly, the combination of both the relational and item-specific processing in the collaborative retrieval led to enhanced recall on the final free-recall test (relies on relational processing) and final recognition test (relies on item-specific processing). Specifically, participants who previously engaged in collaborative retrieval had higher clustering and category access (indicating relational processing), and items recalled per category and recognition score (indicating item-specific processing).

Taken together these studies show that testing promotes both the relational and item-specific processing, resulting in the enhanced final test performance.

**Linking Relational and Item-specific Processing with Testing with Lures**

Research has shown that the intervening free-recall test promotes both the relational and item-specific processing, resulting in the positive testing effect (e.g., Rawson, Wissman & Vaughn, 2015; Wissman & Rawson, 2015, Zaromb & Roediger, 2010). Given that intervening multiple-choice testing in the standard format (4 alternatives) results in the positive testing effect (e.g., Marsh, Bjork, & Bjork, 2006), it is plausible that multiple-choice testing promotes similar processing. However, when the numbers of lures on the intervening test increases, it disrupts the relational and item-specific processing. As the number of lures increases, the test becomes more difficult and students may take a longer amount of time to discriminate the correct answer from the lures instead of a quick scanning. For example, while reading through the alternatives for the category animal, students might engage in relational processing to associate the alternatives in the intervening test as exemplars belonging to animal. As a result, the new category-exemplar organization (which includes the lures) interferes with the original organization during encoding,

disrupting the relational processing. Also, item-specific processing might be operating during the discrimination process. Because item-specific processing draws attention to the individual alternatives, the correct answer will lose its distinctiveness, thereby disrupting the item-specific processing for the correct answer. But, the original organization should be intact for the restudy condition which is not exposed to lures. As a result, the positive testing effect is attenuated or negated in the final free-recall test that relies on relational processing, or the recognition and cued-recall test that relies on item-specific processing (e.g., Jang, Pashler & Huber, 2014; Odegard & Koen, 2007). That is, when items belonging to a taxanomic category are included on a multiple-choice test, both the final free-recall test and the recognition and cued recall tests will show decreased performances.

In contrast, the relational and item-specific processing should be intact when the intervening test does not contain lures such as the correct *all-of-the-above* question. Because there are no lures to interfere with the processing, the relational and item-specific processing will be less disrupted. As result, on a final free-recall that is reliant on the relational processing and the final recognition test that relies on the item-specific processing, testing results in higher recall (e.g., Bishara & Lanzo, 2015).

The purpose of the current experiments is to examine whether multiple-choice retrieval practice with additional lures disrupts the relational and item-specific processing, resulting in a decreased positive testing effect and increased negative side effects (i.e., lure intrusion). In experiment 1 the relational and item-specific processing was compared in a non-inclusive intervening multiple-choice test, whereas experiment 2 was compared these processing in an inclusive multiple-choice test.

**Experiment 1**

In experiment 1, participants studied items belonging to six different taxonomic categories. They took an intervening multiple-choice test and a final free-recall test for all the categories after a 5 minute delay. The number of lures on the intervening test was manipulated with 1 or 5 lures.

***Hypothesis 1:***

It is hypothesized that there would be a positive testing effect. That is, the final test proportion correct for the tested questions would be higher than the control questions. However, the magnitude of the positive testing effect would be attenuated as the number of lures on the intervening test increases. That is, the final test accuracy for the tested questions with five lures would be significantly lower than one lure questions, and this is would be due to the disruption of the relational and item-specific processing.

***Hypothesis 2:***

It is hypothesized there would be a negative side effect of testing. That is, the lure intrusion in the final test would be significantly higher for the tested questions compared to the control questions, and the magnitude of the negative side effect of testing would increase as the number of lures on the intervening test increases and this would be due to the disruption of the relational and item-specific processing.

***Hypothesis 3:***

It is hypothesized that the relational and item-specific processing would be disrupted as the number of lures on the intervening test increases. If relational and item-specific processing is disrupted, the final free recall accuracy should be lower for the information related to the questions with 5 lures than questions with one lure.

## Experiment 1 Method

### Participants and design

116 participants (17.24% males, 78.44% females, 0.04% unspecified) from Auburn University were recruited in exchange for course credit. The learning condition was manipulated between subjects and has 3 levels: restudy, 1 lure test questions, 5 lures test questions (see Figure 1).

### Materials

108 exemplars from 6 different categories were taken from the category norms of Van Overschelde, Rawson, and Dunlosky (2004; See appendix). The 10 most frequently occurring exemplars from each category were eliminated to prevent guessing highly associated items one the final test. For each category, 6 exemplars were randomly chosen to serve as targets. The remaining exemplars served as lures. Because the category norms did not include enough items to have unique lures for all of the questions on the intervening tests, additional lures were created for each category. The additional lures were semantically related to the exemplars in a given category and ranged from 3 to 7 letters matching the exemplars from the norm list. Thus, the lures consist of exemplars from the norm list and the exemplars created by the experimenter for the corresponding category. The stimuli consist of 216 exemplars (36 targets and 180 lures) in total. The categorized exemplars were selected to compute the categorical clustering scores using the Adjusted Ratio of Clustering formula (ARC; Roenker, Thompson, & Brown, 1971). The ARC measure captures the categorical clustering (recall of categorically related words) that corresponds to the categorized studied list and is calculated by measuring the degree to which participants organize related words belonging to the same category from the study list.

The intervening multiple-choice test was divided into two lure types with 36 questions in each condition: 1 lure (2 alternatives) and 5 lures (6 alternatives). For example, “animal -?” had

”giraffe; *lizard* (lure)” in the 1 lure condition and “giraffe; *sheep, monkey, wolf, lizard, jaguar* (lures)” for the 5 lures condition. Note that in the study condition, all target exemplars belonging to a category were presented together, whereas each target exemplar plus the lures (either 1 or 5 lures) was presented one at a time blocked by category in the test conditions. That is, there were six multiple-choice questions per category with different set of lures for each target belonging to a category (see figure 1). The between subject manipulation was chosen instead of the typical within subject to prevent contamination of the lures in the restudy condition and to parse out the processing for each learning condition.

**Procedure**

This experiment consists of three phases (study, learning, recall).

*Study phase*. Each category cue and its 6 targets (e.g., animal – giraffe, rabbit, goat, cheetah, donkey, turtle) was presented for 24 seconds.

*Distractor phase*. Participants solved a math task for 3 minutes.

*Intervening phase*. Participants in the restudy condition once again read the category exemplar targets for 60 seconds each, whereas participants in the testing condition took a multiple-choice test. Based on the test condition, participants were presented with a categorical cue followed by either 2 alternatives with 1 lure or 6 alternatives with 5 lures and was instructed to choose one correct answer. Each question was presented for 10 seconds. The presentation of the cue-target was blocked by taxonomic categories to ensure that the consecutive questions belong to the same category to parallel the cue-target order in the restudy condition. Thus, similar to the restudy condition, in the testing condition participants spent a total of 60 seconds to retrieve the 6 targets belonging to the same category, with 10 seconds per cue-target.

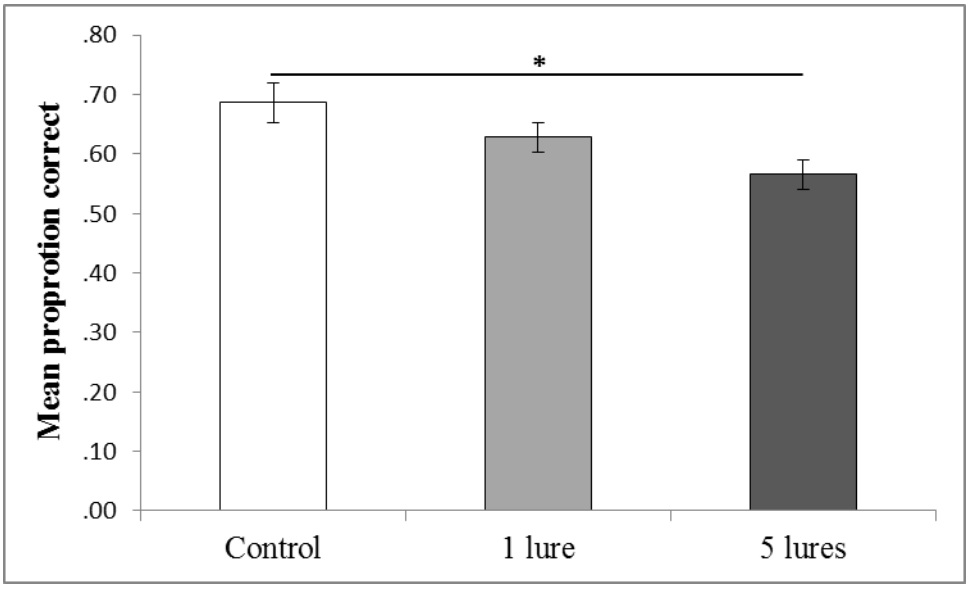
*Distractor phase*. Participants solved a math task for 5 minutes.

*Recall phase.* Participants were randomly assigned to take a free-recall test. Participants were instructed to recall as many targets they could from the study list at the beginning of the experiment and were warned against guessing. The final recall phase was self-paced.

## Experiment 1 Results

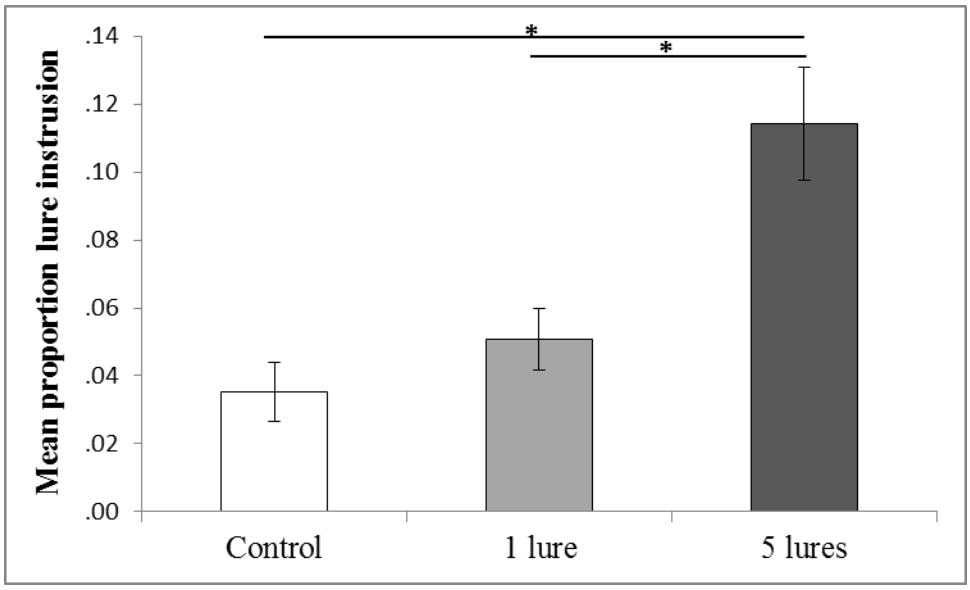
**Final test proportion correct.** Figure 2 displays mean proportion correct on the final free recall test. A one way ANOVA showed a significant main effect, $F(2,115) = 4.62$, $MSE = .14$, $p = .01$, $\eta p^2 = .08$. Contrary to hypothesis, performance was not significantly different on the final test for initially tested 1 lure test questions ($M = .63$, $SD = 1.5$) compared to control questions ($M = .69$, $SD = .21$), $t(113) = 1.48$, $p = .14$ $d = .28$ , indicating the null testing effect. Also, performance was significantly lower on the final test for initially tested 5 lure test questions ($M = .57$, $SD = .16$) compared to control questions ($M = .69$, $SD = .21$), $t(113) = 3.04$, $p = .003$, $d = .57$, indicating the negative testing effect. Against prediction, there was not a difference on the final test for the initially tested 1 lure test ($M = .63$, $SD = 1.5$) and 5 lures test ($M = .57$, $SD = .16$) questions, $t(113) = 1.57$, $p = .12$, $d = .30$.

*Figure 2.* Final test proportion correct



**Lure intrusion.** Figure 3 shows mean intrusion of lure on the final free recall test. A one way ANOVA revealed a significant main effect, $F(2, 115) = 12.03$, $MSE = .07$, $p < .001$, $\eta p^2 = .18$. Supporting hypothesis 2, compared to control questions ($M = .04$, $SD = .05$), participants responded with incorrect lures significantly higher for questions initially tested with 5 lures ($M = .11$, $SD = .10$), $t(113) = 4.62$, $p < .001$ $d = .87$. Also, there was significantly higher intrusion of lures for questions initially tested with 5 lures ($M = .11$, $SD = .10$) compared to 1 lure ($M = .06$, $SD = .01$ ), $t(113) = 3.73$, $p < .001$ $d = .70$. However, there was no significant difference in lure intrusion for control questions ($M = .04$, $SD = .05$) and 1 lure test questions ($M = .06$, $SD = .01$), $t(113) = .91$, $p < .37$ $d = .17$.

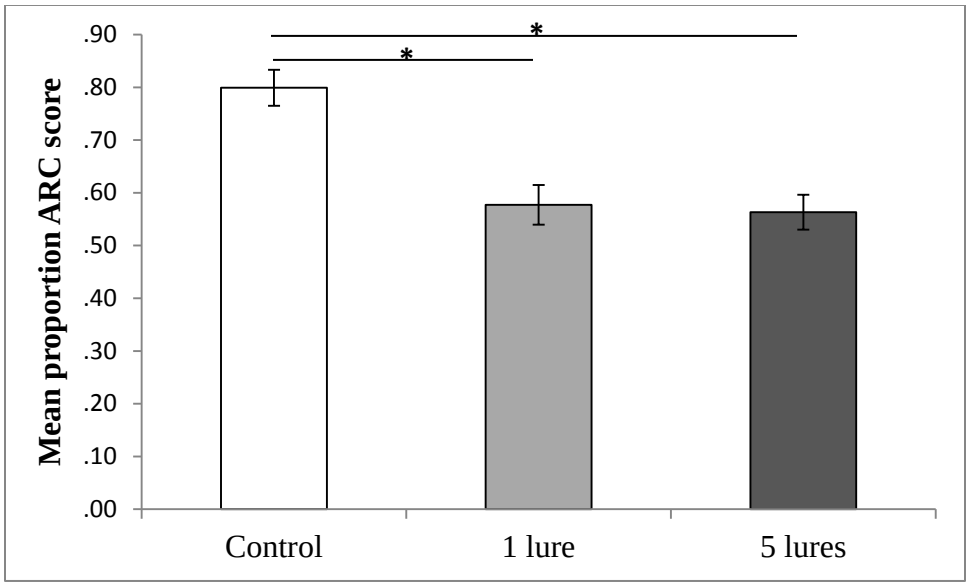
*Figure 3.* Intrusion of lures




**Measures of relational processing.** Figure 4 shows ARC scores. Clustering was computed using the Adjusted Ratio of Clustering formula (Roenker, Thompson, & Brown, 1971). A one way ANOVA showed a significant main effect in ARC score, $F(2,115) = 14.17$, $MSE = .67$, $p = <.001$, $\eta p^2 = .20$. As shown in Figure 3, the ARC score was higher for participants who answered the control questions ($M = .80$, $SD = .21$) compared to those who answered 1 lure test

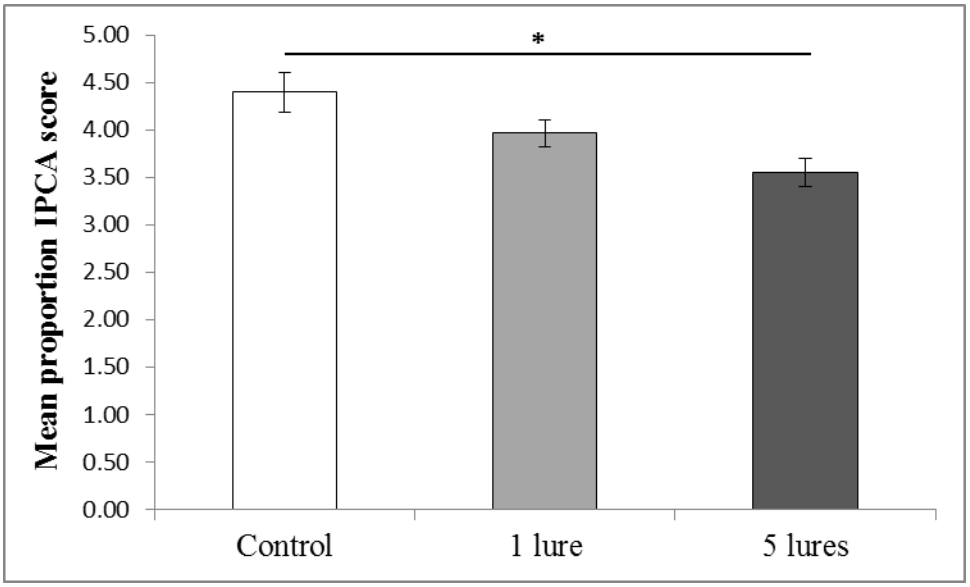
questions (*M* = .58, *SD* = .23), *t*(113) = 4.48, *p* < .001 *d* = .84 and 5 lures test questions (*M* = .56, *SD* = .21),*t*(113) = 4.75, *p* < .001, *d* = .89. Against prediction, the ARC score between the 1 lure test questions (*M* = .58, *SD* = .23) and 5 lures test questions (*M* = .56, *SD* = .21) was insignificant *t*(113) = .28, *p* = .78 *d* = .05.

*Figure 4.* ARC score




**Measures of item-specific processing.** Figure 5 shows IPCA scores. A one way ANOVA showed a significant main effect in IPCA score, *F*(2,115) = 6.78, *MSE* = .14, *p* = .003, $\eta p^2$ = .01. Posthoc t-test revealed that IPCA score was significantly lower in the 5 lures test questions (*M* = 3.56, *SD* = .92) compared to control questions (*M* = 4.40, *SD* = 1.28), *t*(113) = 3.52, *p* = .001 *d* = .66. Contrary to prediction, IPCA score was not significantly different in control questions (4.40) compared to 1 lure test questions (*M* = 3.97, *SD* = .90), *t*(113) = 1.80, *p* = .07, *d* = .34, nor between 1 lure test questions (*M* = 3.97, *SD* = .90) compared to 5 lures test question (*M* = 3.56, *SD* = .92), *t*(113) = 1.73, *p* = .09, *d* = .33.

*Figure 5.* IPCA score




## Experiment 1 Discussion

Experiment 1 tested the prediction that the presence of additional lures on the 5 lures questions compared to 1 lure questions would disrupt both the relational and item-specific processing. Against hypothesis, the results revealed no evidence of the positive testing effect. Overall, participants who reread the study materials performed better on the final free-recall test compared to those who took the intervening multiple-choice, indicating the null testing effect for 1 lure questions and the negative testing effect for 5 lures questions. This result failed to replicate previous findings (Roediger & Marsh, 2005) showing that practice testing with multiple-choice with additional lures increases retention on the final test compared to the control questions, resulting in the positive testing effect. However, increasing the number of lures increased the intrusion of lures on the final free-recall test. Importantly, the disruption of the relational processing corresponds to the null testing effect found in testing with 1 lure, whereas disruption of both the relational and item-specific processing corresponds to the negative testing effect found in testing with 5 lures. Furthermore, both the processing is disrupted to the same extent for

both the questions, and that explains the lack of difference in the magnitude of the negative testing effect. In summary, taking an intervening multiple-choice test with 5 lures disrupted the initial intact relational and item-specific processing thereby reducing retention on the final free-recall test that relies on both this processing.

**Experiment 2**

Experiment 1 was designed to examine whether additional lures on a non-inclusive intervening multiple-choice test will disrupt the relational and item-specific processing. Experiment 2 further addressed this question by examining the presence of lures on an inclusive intervening multiple-choice test with "*none-of-the-above*" option as the correct alternative or incorrect alternative. Prior studies show that the positive testing effect is negated when previously tested with the correct ''*none-of-the-above*'' (all alternatives are lures), but present for the incorrect ''none-of-the-above'' (target present among the incorrect lures) (Jang, Pashler & Huber, 2014; Odegard & Koen, 2007). In contrast to the presence of only lures on the correct ''*none-of-the-above*'', absence of lures in a correct "*all-of-the-above*" enhances the positive testing effect (Bishara & Lanzo, 2015; Paneerselvam & Callender, 2016). Thus, similar to non-inclusive test, it is plausible that the relational processing and item-specific processing is disrupted in the inclusive test questions that have only the lures. That is, the presence of only lures in the correct *none-of-the-above* (all lures) questions disrupts the relational and item-specific processing.

***Hypothesis 1:***

It is hypothesized that there would be a negative testing effect for the correct none-of-the-above questions. That is the final test proportion correct for the intervening question with the correct NOTA would be significantly lower than the control questions. However, there will be a positive testing effect or null testing effect for the wrong none-of-the-

above questions. That is the final test proportion correct for the intervening question with the wrong NOTA will be significantly higher or same as the control questions, and this is would be due to the disruption of the relational and item-specific processing.

***Hypothesis 2:***

It is hypothesized there would be a negative side effect of testing. That is, the lure intrusion in the final test would be significantly higher for the tested questions compared to the control questions, and the lure intrusion would be significantly higher for the correct NOTA than the incorrect NOTA, and this is would be due to the disruption of the relational and item-specific processing.

***Hypothesis 3:***

It is hypothesized that the relational and item-specific processing will be significantly disrupted for the correct-NOTA questions than the wrong-NOTA and control questions. If relational and item-specific processing is disrupted, the final test accuracy should be significantly lower for the correct-NOTA questions than control questions and wrong-NOTA questions on the final free-recall test that relies on relational and item-specific processing. Further, the magnitude of the disruption of the relational processing will be higher for the correct-NOTA (all lures) than the wrong-NOTA.

## Experiment 2 Method

### Participants and design

108 participants (23 % male, 77% female) from Auburn University were recruited in exchange for course credit. The learning condition was manipulated between subjects and has 3 levels: restudy, correct-NOTA, wrong-NOTA (see Figure 6).

### Materials

The materials were identical to those of experiment 1 except for the intervening multiple-choice test. The intervening multiple-choice test was divided into two lure types with 36 questions in each condition: correct-NOTA (all lures) and incorrect-NOTA (target present along with lures). To prevent students from guessing the correct answers, the type of questions was mixed up in each test condition. The correct-NOTA condition had 5 questions with NOTA as the correct answer and 1 question with NOTA as the incorrect answer whereas the wrong-NOTA condition had 5 questions with NOTA as the incorrect answer and 1 question with NOTA as the correct answer.

**Procedure**

The procedure was identical to those of experiment 1 except for the learning phase.

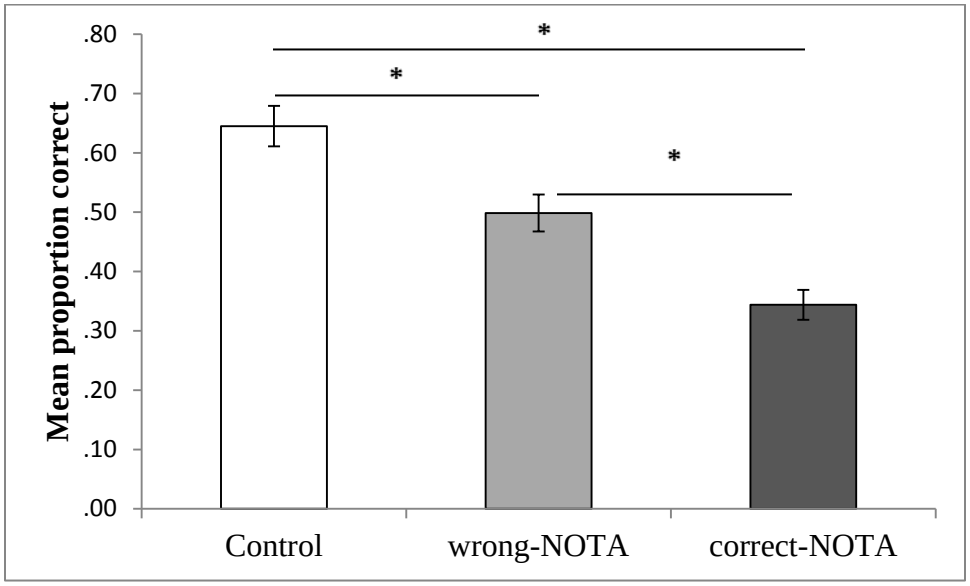
*Learning phase*. Participants in the restudy condition once again read the category exemplar targets for 60 seconds each, whereas participants in the testing condition took a multiple-choice test. Based on the test condition, participants were presented with a categorical cue followed by either correct-NOTA or wrong-NOTA questions and was instructed to choose one correct answer. Each question was presented for 10 seconds. The presentation of the cue-target was blocked by taxonomic categories to ensure that the consecutive questions belong to the same category to parallel the cue-target order in the restudy condition. Thus, similar to the restudy condition, in the testing condition participants spent a total of 60 seconds to retrieve the 6 targets belonging to the same category, with 10 seconds per cue-target.

## Experiment 2 Results

**Final test proportion correct.** Figure 7 displays mean proportion correct on the final free recall test. A one way ANOVA showed a significant main effect, $F(2, 107) = 24.57$, $MSE = .82$, $p < .001$, $\eta p^2 = .32$. Supporting hypothesis 1, performance was significantly lower on the final test

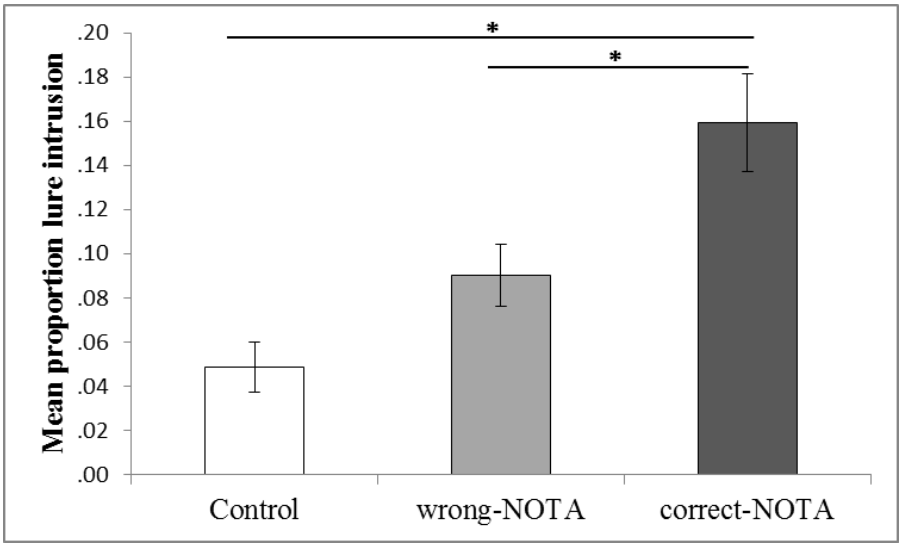
for the initially tested correct-NOTA questions (*M* = .34, *SD* = 1.5) than the control questions (*M* =.65, *SD* = .20), *t*(105) = 7.01, *p* = <.001, *d* = 1.37, indicating the negative testing effect. Against prediction, performance was significantly lower on the final test for initially tested wrong-NOTA questions (*M* = .50, *SD* = .19) than the control questions (*M* = .65, *SD* = .20), *t*(105) = 3.41, *p* = .001, *d* = .67, indicating the negative testing effect. Furthermore, performance on the final test was significantly higher for the initially tested wrong NOTA (*M* = .50, *SD* = .19) than the correct NOTA (*M* = .34, *SD* = 1.5), *t*(105) = 3.60, *p* < .001, *d* = .70.

*Figure 7.* Final test proportion correct



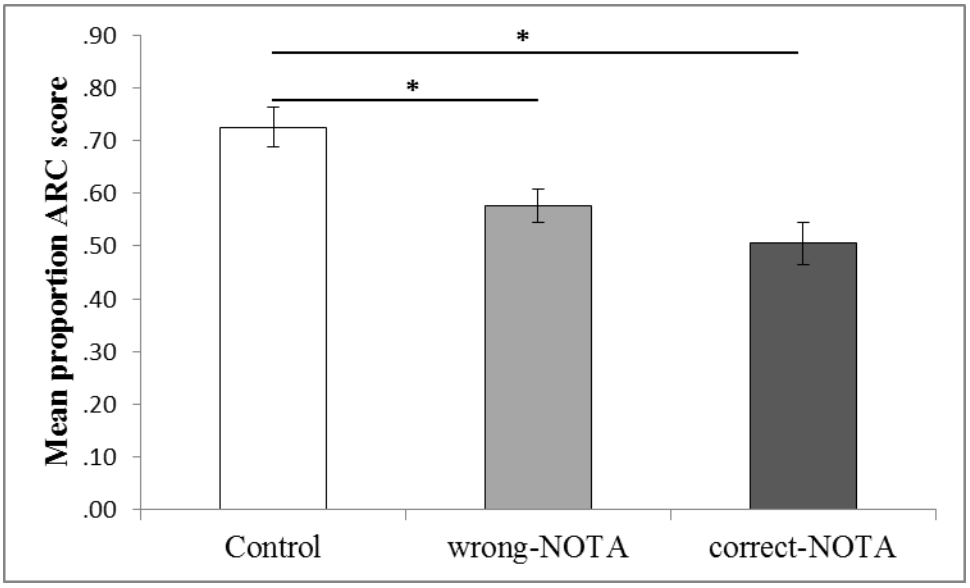

**Lure intrusion.** Figure 8 shows mean intrusion of lure on the final free recall test. A one way ANOVA revealed a significant main effect, *F*(2, 107) = 11.51, *MSE* = .11, *p* < .001, $\eta p^2$ = .18. Compared to control questions (*M* = .05, *SD* = .07), the intrusion of lures is significantly higher for the correct NOTA (*M* = .16, *SD* = .13), *t*(105) = 4.75, *p* < .001, *d* = .93 and wrong NOTA (*M* =.09, *SD* = .08), *t*(105) = 1.79, *p* = .08, *d* = .35. Additionally, the intrusion of lures for the correct NOTA (*M* = .16, *SD* = .13) was significantly higher than wrong NOTA (*M* = .09, *SD* = .08), *t*(105) = 2.96, *p* = .004, *d* = .58. These results confirm hypothesis 2.

*Figure 8.* Lure intrusion



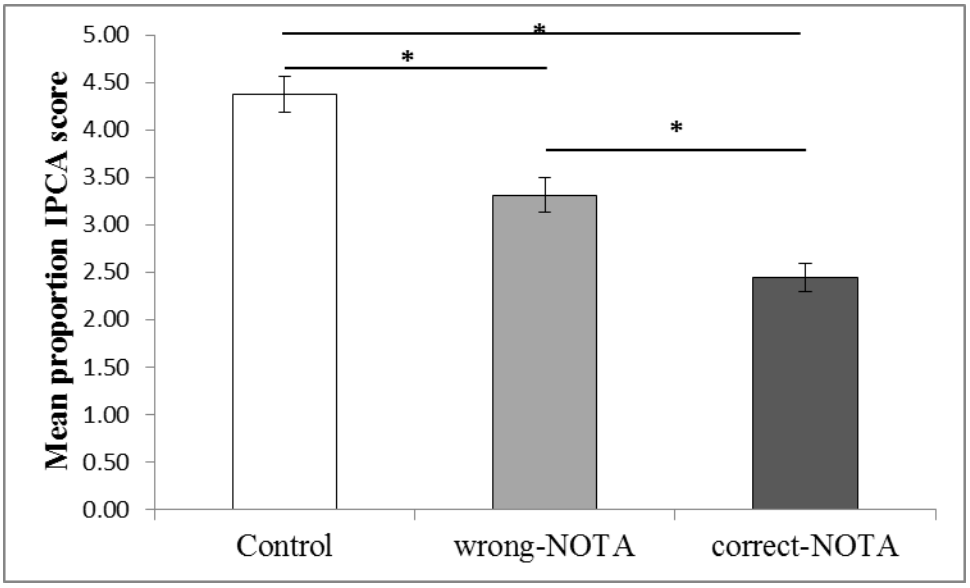

**Measures of relational processing.** Figure 9 shows mean ARC score on the final free recall test. A one way ANOVA showed a significant main effect in ARC score, $F(2, 107) = 9.35$, $MSE = .45$, $p < .001$, $\eta p^2 = .15$. Against prediction 3, compared to the control question ($M = .73$, $SD = .23$), the ARC score was significantly lower for the correct-NOTA ($M = .51$, $SD = .24$), $t(105) = 4.23$, $p < .001$, $d = .83$ and the wrong-NOTA ($M = .58$, $SD = .19$), $t(105) = 2.89$, $p = .005$, $d = .56$. However, against prediction, there was no significant difference between the correct NOTA ($M = .51$, $SD = .24$) and wrong NOTA ($M = .58$, $SD = .19$), $t(105) = 1.35$, $p = .18$, $d = .26$.

*Figure 9.* ARC Score




**Measures of item-specific processing.** Figure 10 shows IPCA scores. A one way ANOVA showed a significant main effect, $F(2, 107) = 31.23$, $MSE = 33.36$, $p < .001$, $\eta p^2 = .37$. Compared to the control question ($M = 4.37$, $SD = 1.12$), IPCA score was significantly lower for correct NOTA ($M = 2.45$, $SD = .87$), $t(105) = 7.89$, $p < .001$, $d = 1.54$ and wrong NOTA ($M = 3.31$, $SD = 1.09$), $t(105) = 4.34$, $p < .001$, $d = .85$. Also, the IPCA score was significantly lower for the correct NOTA ($M = 2.45$, $SD = .87$) than wrong NOTA ($M = 3.31$, $SD = 1.09$), $t(105) = 3.55$, $p = .001$, $d = .69$. This finding supports hypothesis 3.

*Figure 10.* IPCA Score




**Experiment 2 Discussion**

Experiment 2 examined the prediction that the presence of only lures in the correct *none-of-the-above* questions disrupts the relational and item-specific processing, resulting in decreased retention on the final test. Consistent with the hypothesis, multiple-choice testing with the correct *none-of-the-above* questions disrupted relational and item-specific processing, resulting in the negative testing effect. The finding that participants performed poorly on the final test and produced more lures as answers on the final test after answering the correct *none-of-the-above* aligns well with Odegard and Koen (2007) study. However, the negative testing effect was again observed for the wrong *none-of-the-above,* while previous study (Odegard & Koen, 2007) observed the positive testing effect. Interestingly, the negative testing effect for the wrong *none-of-the-above* is once again explained by disruption of the relational and item-specific processing. Furthermore, the disruption of the item-specific processing rather than relational processing

explains the magnitude of the negative testing effect for the correct *none-of-the-above* and wrong *none-of-the-above* questions.

## General Discussion

Despite research showing multiple-choice testing with additional lures results in reduced retention and increased production of lures on a later test (Odegard & Koen, 2007; Roediger & Marsh, 2005), the basic underlying mechanisms of the negative outcomes remain poorly understood. Past research has shown that taking an initial free-recall test (e.g., Wissman & Rawson, 2015) and cued-recall test (e.g., Rawson, Wissman & Vaughn, 2015) promote both the relational and item-specific processing resulting in the positive testing effect. Therefore, it is possible that taking an initial multiple-choice test with additional lures might disrupt the both the relational and item-specific processing, resulting in decreased retention on the final test. The goal of this study was to examine whether the type of processing: relational versus item-specific, might hold possibilities in explaining the reduced retention associated with multiple-choice testing with additional lures. Across two experiments, results support this prediction.

Experiment 1 found that taking an initial multiple-choice question with 5 lures results in the negative testing effect and this finding is due to disrupted relational and item-specific processing, whereas the null testing effect found with 1 lure questions is a result of disrupted relational processing. Experiment 2 found that taking initial multiple-choice questions with correct-NOTA (only lures) and wrong-NOTA (correct answer present among the lures) disrupted both the relational and item-specific processing resulting in the negative testing effect.

Past research have shown that successful recall is dependent on the final test formats that afford similar processing as the learning method. For instance, in Zaromb and Roediger (2010) study, the final free-recall test that is reliant on the relational processing and item-specific

processing benefitted from the free-recall practice test that promoted both these processing. Conversely, disruption of both these processing could result in decreased retention on the final free-recall test that relies on these processing. In line with this reasoning, the current study shows that initial multiple-choice testing with additional lures disrupts both relational and item-specific processing, resulting in decreased retention on the final free-recall test that relies on these processing.

Surprisingly, however, there was no positive testing across both the experiments, as the rereading control group performed better than the testing group on the final test. This failed to confirm the hypothesis and failed to echo previous research showing an overall positive testing effect for the initially tested questions with additional lures (1 lure vs 5 lures) in a non-inclusive format (e.g., Roediger & Marsh, 2005) and the wrong none-of-the-above questions (target present among lures) in an inclusive format (e.g., Odegard & Koen, 2007). However, the null testing effect and the negative testing effect found in these questions relate to the disrupted relational and item-specific processing. Thus, the relational and item-specific processing theoretical framework can explain not only the negative testing effect found in additional lures (i.e. 5 lures; correct none-of-the-above), but also the overall reduced retention in questions without additional lures.

The absence of the positive testing effect could be due to several reasons. First, it is difficult to obtain the positive testing effect with a proper restudy control, in which the final test performance for the information that is tested on the intervening phase is compared to those presented for restudy for an equal duration (e.g.,Carpenter & DeLosh, 2006). In the earlier multiple-choice studies, the positive testing effect was obtained when the tested information on the intervening test is contrasted with information that is not tested or restudied on the

intervening phase (i.e., control questions) (e.g., Odegard & Koen; Roediger & Marsh, 2005). Thus, the absence of the positive testing effect reported here is possibly due to the proper restudy control used. For instance, with the proper restudy control, participants reread the categorical cue and its entire targets at a time both in the initial study phase and the intervening phase. Such repeated exposure to the study items may have provided ample time for the participants to strengthen their relational and item-specific processing. In contrast, for the testing group, each categorical cue and its corresponding single target is presented one at a time in the intervening phase. Therefore, participants in the testing group engaged in item-specific processing more than relational to discriminate the answer options. In the process, they incorporated the lures in their memory which might have interfered with their original intact processing. Given that multiple-choice testing naturally affords item-specific processing, participants had limited opportunity to engage in relational processing which explains the disrupted relational processing in the additional lures group. Future investigation will be needed to examine whether providing correct answer feedback could combat interference of lures thus strengthening the item-specific processing. Previous multiple-choice testing in the standard format has observed a decrease in the negative testing effect when provided with feedback (e.g., Butler & Roediger, 2008).

A second potential possibility for the negative testing effect could be that the manipulations of the categorical cue and target organization. The study items were presented in a blocked fashion (categorical cue and all the targets belonging to that categorical cue) without randomization in the study phase and the intervening phase. Compared to the testing group which had exposure to a single target belonging to a categorical cue for 10 seconds each in the intervening phase, the study group had the entire 60 seconds to view the categories and its corresponding targets. Such blocked presentations of the study materials could be an added

advantage to the study group. However, when the study items were randomized and not presented in a blocked fashion (i.e., a category and a target belonging to that category, with the limitation that no target from the same category appears one after the other), there was no negative testing effect but the null testing effect (equal performances in the study and testing conditions) (See Appendix B).

Third, it is also possible that the absence of the positive testing effect may be due to study material used. Whereas this study used categorized word list, previous research on testing with additional lures used prose passages (e.g., Odegard & Koen, 2007; Roediger & Marsh, 2005). Although Jang, Pashler, and Huber (2014) obtained the positive testing effect with the word list, they did not include a proper restudy control. Third, the short retention interval between the learning phase and the final recall could have led to the absence of the positive testing effect. These finding is in line with studies showing restudying is beneficial than testing at short retention interval (e.g., Wheeler, Ewers, & Buonanno, 2003).

In summary, the current studies examined the relational and item-specific processing, a general mechanism that has been applied to other learning techniques and describing its effects on multiple-choice testing. The results of the present study demonstrate that taking an initial multiple-choice testing with additional lures disrupts both the relational and item-specific processing, resulting in reduced retention as evidenced by the reduced ARC and IPCA scores. However, repeated studying results in robust relational and item-specific processing as indicated by the increased ARC and IPCA scores. The data also highlight that the relational and item-specific processing theoretical framework explains the underlying negative outcomes of multiple-choice testing in general. Moving forward, it will be important to conduct additional research by modifying the experimental designs to clarify the atypical negative testing effect

found in questions with no additional lures, as well as reverse the negative testing effects found in these questions. As a general conclusion then, the item-specific and relational theoretical framework appears to be a promising theory to support the negative outcomes of the multiple-choice testing.

**References**

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940.

Bishara, A. J., & Lanzo, L. A. (2015). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*(7), 1013-1028.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings.

Bjork, R. A. (1975). 5 Retrieval as a memory modifier: an interpretation of negative recency and related phenomena.

Brown, A. S. (1988). Experiencing misspellings and spelling performance: why wrong isn't right. *Journal of Educational Psychology, 80*, 488–494.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941-956.

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527.

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

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*(6), 760-771.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553.

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & cognition*, *40*(4), 528-539.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology*, *58*(1), 17.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 588.

Grimaldi, P. J., Poston, L., & Karpicke, J. D. (2015). How does creating a concept map affect item-specific encoding?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1049.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107-112.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562-567.

Huelser, B. J., & Marsh, E. J. (2006). Does guessing on a multiple-choice test affect later cued recall. In *Poster presented at the Annual Meeting of the Psychonomic Society, Houston, TX*.

Hunt, R. (2012). *Basic growth analysis: plant growth analysis for beginners*. Springer Science & Business Media.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 497-514.

Hunt, R. R., & Elliot, J. M. (1980). The role of nonsemantic information in memory: Orthographic distinctiveness effects on retention. *Journal of Experimental Psychology: General*, *109*(1), 49.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *32*(4), 421-445.

Hunt, R. R., & Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 81.

Hunt, R. R., & Worthen, J. B. (Eds.). (2006). *Distinctiveness and memory*. Oxford University Press.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of verbal learning and verbal behavior*, *17*(6), 649-667.

Jang, Y., Pashler, H., & Huber, D. E. (2014). Manipulations of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology*, *106*(2), 435.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning

text materials?. *Memory & Cognition*, *38*(1), 116-124.

Knouse, L. E., Rawson, K. A., Vaughn, K. E., & Dunlosky, J. (2016). Does Testing Improve Learning for College Students With Attention-Deficit/Hyperactivity Disorder?. *Clinical Psychological Science*, *4*(1), 136-143.

Lipowski, S. L., Pyc, M. A., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. *Developmental psychology*, *50*(4), 994.

Little, J. L., & Bjork, E. L. (2010). Multiple-choice testing can improve the retention of non-tested related information. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1535-1540). Cognitive Science Society Austin, TX.

Mandler, G. (1967). Organization and memory.

Mandler, G. (1972). Organization and recognition.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194-199.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513.

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 54.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*(2), 192-201.

Meyer, A. N., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and aging*, *28*(1), 142.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

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

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. *The psychology of learning and motivation: Advances in research and theory*, *45*, 175-214.

Nungester, R. J. & Duchastel, p. C. (1982). Testing versus review: Efefcts on retention. *Journal of Applied Psychology, 74*(1), 18-22.

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15*(8), 873-885.

Paneerselvam, B., & Callender, A.A. (2016). Benefits of all-of-the-above questions on multiple-choice test. Manuscript in preparation.

Peterson, D. J., & Mulligan, N. W. (2012). A negative effect of repetition in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1786.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287.

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, *1*(1), 19-40.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language*, *60*(4), 437-447.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological review*, *88*(2), 93.

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.

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology*, *17*(1), 52.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155.

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

Rundus, D. (1973). Negative effects of using list items as recall cues. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 43-50.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, *3*(4), 207-217.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice,

and hybrid tests. *Memory*, *22*(7), 784-802.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641.

Toppino, T. C., & Ann Brochin, H. (1989). Learning from tests: The case of true-false examinations. *The Journal of Educational Research*, *83*(2), 119-124.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research*, *86*(6), 357-362.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289-335.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240-245.

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 127.

Wissman, K. T., & Rawson, K. A. (2015). Why does collaborative retrieval improve memory? Enhanced relational and item-specific processing. *Journal of Memory and Language*, *84*, 75-87.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995-1008.

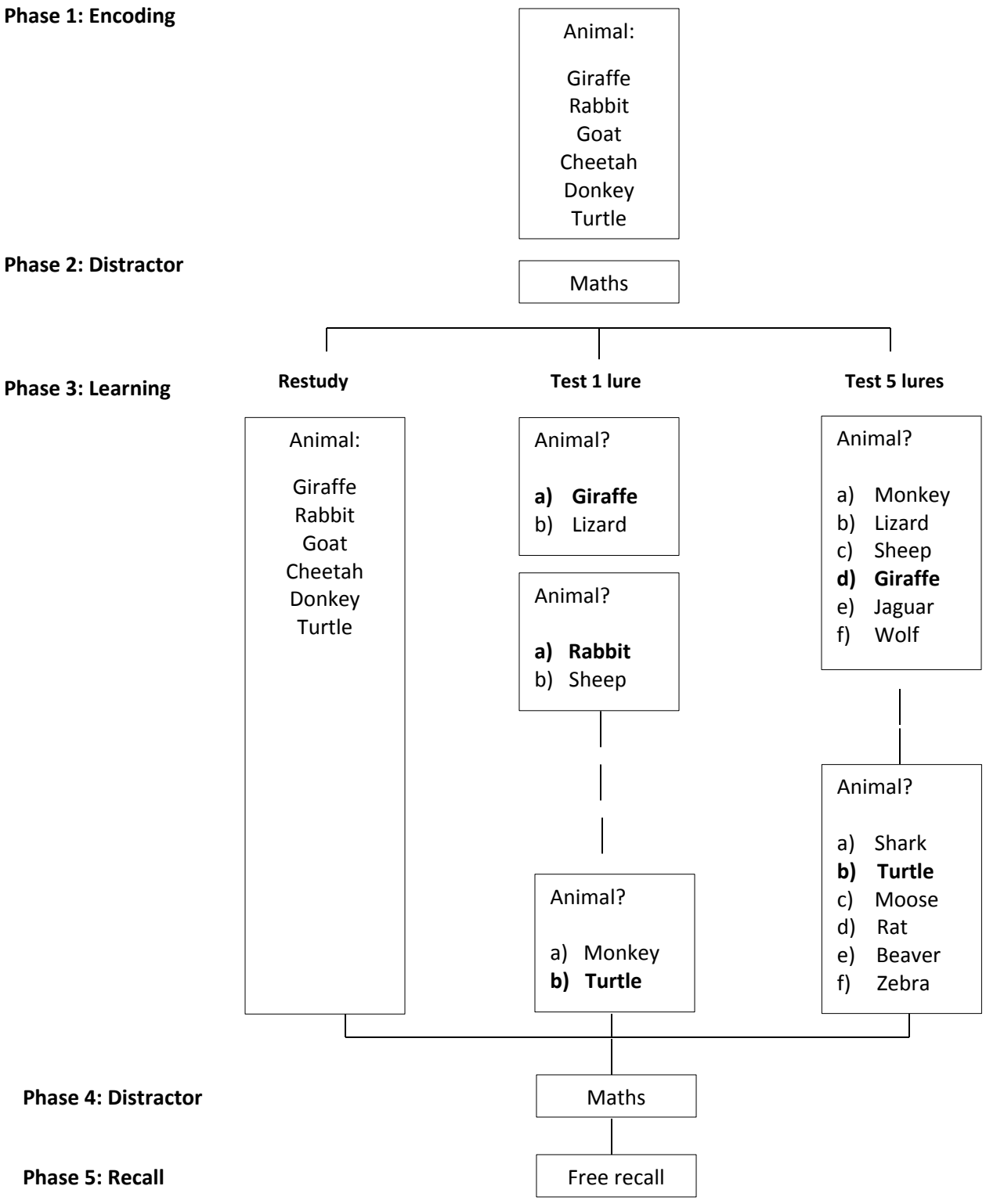



Figure 1. Experimental design and procedure for experiment 1. The correct answers (targets) are in bold.

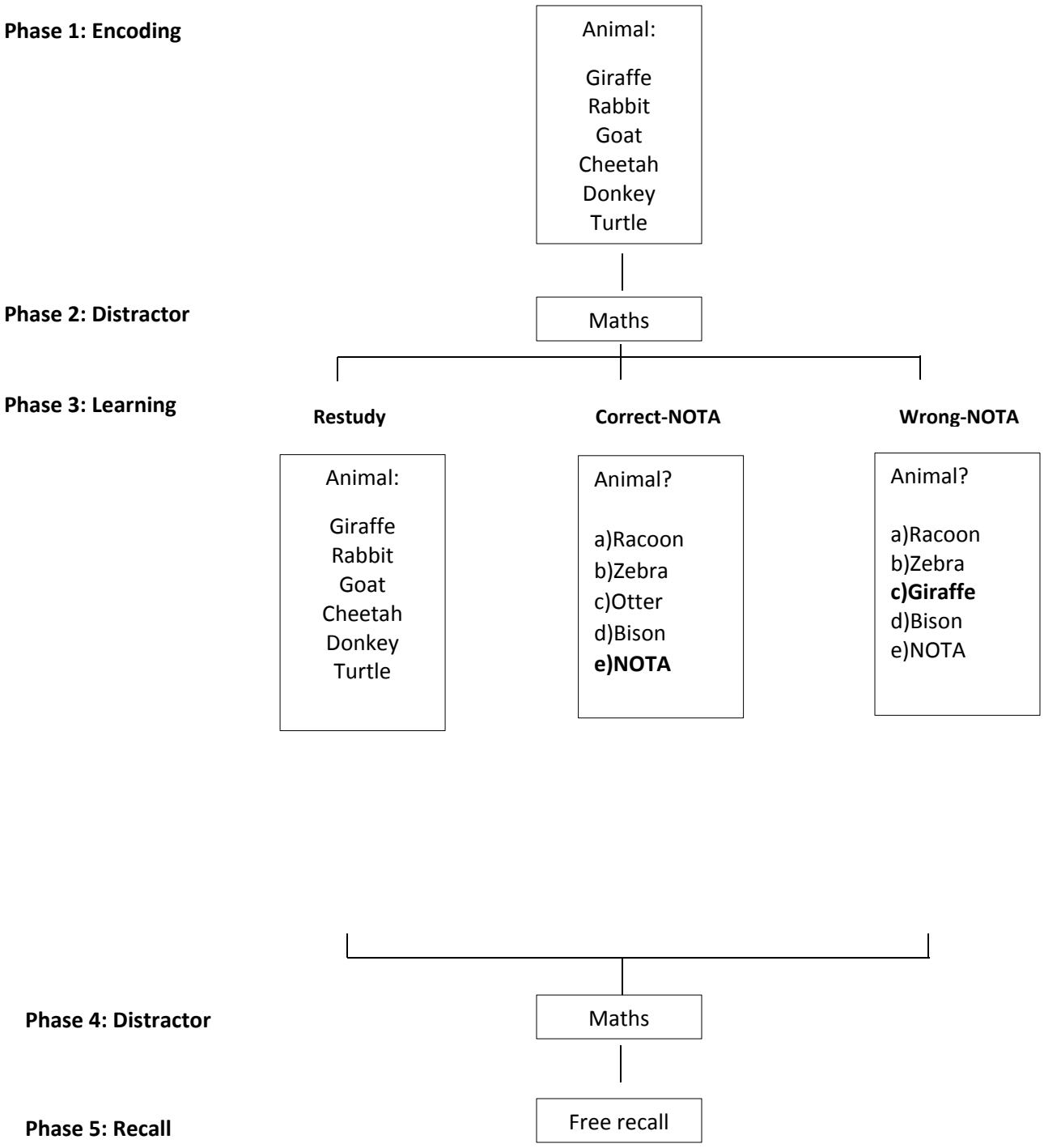



Figure 6. Experimental design and procedure for experiment 2. The correct answers (targets) are in bold.

Appendix A

Items used in Experiments 1 and 2

| Category cue | Target | Lures | Lures (created by experimenter) |
|---|---|---|---|
| Animal | | | |
| | Giraffe | Pig | Koala |
| | Rabbit | Rat | Elk |
| | Goat | Squirrel | Frog |
| | Cheetah | Zebra | Jaguar |
| | Donkey | Moose | Leopard |
| | Turtle | Sheep | Whale |
| | | Raccoon | Snail |
| | | Wolf | Shark |
| | | Fox | |
| | | Hamster | |
| | | Elk | |
| | | Lizard | |
| Occupation | | | |
| | Manager | Secretary | Actor |
| | Cook | Policeman | Pharmacist |
| | Banker | Athlete | Tailor |
| | Carpenter | Businessman | Driver |
| | Therapist | Janitor | Plumber |
| | Scientist | Student | Psychiatrist |
| | | | Masseur |
| | | | Artist |
| | | | Waiter |
| | | | Masseur |
| | | | Model |
| | | | Judge |
| | | | Musician |
| | | | Principal |
| | | | Butler |
| | | | Pilot |
| | | | Priest |
| | | | Dancer |
| | | | Baker |
| | | | Soldier |
| | | | Midwife |
| | | | Singer |
| | | | Surgeon |

| | | | |
|---|---|---|---|
| | | | Astrologer |
| Fruit | | | |
| | Grapefruit | Tomato | Avocado |
| | Lemon | Plum | Coconut |
| | Raspberry | Mango | Cranberry |
| | Melon | Cherry | Lychee |
| | Papaya | Blueberry | Huckleberry |
| | Apricot | Cantaloupe | Mandarine |
| | | Lime | Pomelo |
| | | Tangerine | Clementine |
| | | Nectarine | Quince |
| | | Honeydew | Raisin |
| | | Star fruit | Olive |
| | | | Blueberry |
| | | | Date |
| | | | Gooseberry |
| | | | Jackfruit |
| | | | Mulberry |
| | | | Lime |
| | | | Guave |
| | | | Fig |
| Body Parts | | | |
| | Neck | Stomach | Eyebrow |
| | Chest | Heart | Cheek |
| | Ankle | Knee | Eyelid |
| | Tongue | Brain | Forearm |
| | Wrist | Hair | Knuckle |
| | Muscle | Elbow | Spine |
| | | Shoulder | Heel |
| | | Back | Sole |
| | | Butt | Chin |
| | | Lip | Toe Nail |
| | | Thigh | Waist |
| | | Face | Thumb |
| | | Liver | |
| | | Lung | |
| | | Teeth | |
| | | Torse | |
| | | Bone | |
| | | Penis | |
| | | Breast | |
| | | Hip | |
| | | Finger Nail | |

| | | | |
|---|---|---|---|
| Clothing | | | |
| | Dress | Skirt | Swimsuit |
| | Gloves | Jeans | Bow |
| | Boxer | Coat | Bandana |
| | Blouse | Tshirt | Cap |
| | Belt | Sweatshirt | Tights |
| | Undershirt | Scarf | Vest |
| | | Tank top | Suit |
| | | Tie | Pyjamans |
| | | Panties | Blazer |
| | | | Slacks |
| | | | Raincoat |
| | | | Cardigan |
| | | | Bathrobe |
| | | | Tuxedo |
| | | | Uniform |
| | | | Suspender |
| | | | Trunks |
| | | | Robes |
| | | | Bikini |
| | | | Nightgown |
| | | | Diaper |
| Bird | | | |
| | Pigeon | Seagull | Quail |
| | Canary | Dove | Peacock |
| | Penguin | Parakeet | Goose |
| | Finch | Falcon | Turkey |
| | Woodpecker | Owl | Kingfisher |
| | Blackbird | Ostrich | Puffin |
| | | Raven | Hornbill |
| | | Duck | Nightingale |
| | | Mockingbird | Toucan |
| | | Flamingo | Mynah |
| | | Oriole | Flamingo |
| | | Chicken | Towee |
| | | Vulture | Pelican |
| | | | Stork |
| | | | Rooster |
| | | | Swan |
| | | | Wagtail |

# Appendix B

Study with methodological difference than in Experiment 1 that showed the null testing effect

## Method

73 participants from Auburn University were recruited in exchange for course credit. The learning condition was manipulated between subjects and has 3 levels: restudy, 1 lure test questions, 5 lures test questions. The materials and procedure were the same as in Experiment 1, except that (1) in the study phase, participants were presented with each categorical cue and a single target for 4 seconds in a randomized order (2) in the intervening phase, participants in the study condition studied each category and its target for 10 seconds in a randomized order. Participants in the test conditions were presented with each categorical cue and answer options for 10 seconds in a randomized order.

## Results and discussion

**Final test proportion correct**. Table 1 displays mean proportion correct on the final free-recall test. There were no significant differences among group means as determined by one-way ANOVA, $F(2,72) = .42$, $MSE = .01$, $p = .66$, $\eta p^2 = .01$.

**Lure intrusion.** Table 1 shows mean intrusion of lure on the final free recall test. There were no significant differences among group means as determined by one-way ANOVA, $F(2,72) = .36$, $MSE = .00$, $p = .70$, $\eta p^2 = .01$.

**Measures of relational processing.** Table 1 shows mean ARC scores. There were no significant differences among group means as determined by one-way ANOVA, $F(2,72) = 1.34$, $MSE = .10$, $p = .27$, $\eta p^2 = .04$.

**Measures of item-specific processing.** Table 1 shows mean IPCA scores. There were no significant differences among group mean as determined by one-way ANOVA, $F(2,72) = .44$, $MSE = .35$, $p = .64$, $\eta p^2 = .01$.

**Table 1.**

Mean proportion correct, lure intrusion, ARC score, and IPCA score on the final test as a function of intervening activity

| | Intervening Activity | | |
|---|---|---|---|
| | Study | Test 1 Lure | Test 5 Lures |
| Proportion Correct | .48 (.18) | .48 (.13) | .44 (.16) |
| Lure Intrusion | .07 (.10) | .06 (.05) | .08 (.11) |
| ARC Score | .39 (.25) | .32 (.27) | .26 (.28) |
| IPCA score | 3.00 (1.01) | 2.93 (.79) | 2.76 (.86) |

*Note.* Standard Deviations are in Parentheses

Participants who reread the study material and those who took the intervening test performed equally on the final test when the category-target was presented in randomized fashion in the intervening phase. This finding corresponds to the relational and item-specific processing which are disrupted to the same extent for participants in all the conditions.