# An Exploration of the Negative Effects of Repetition and Testing on Memory

**S. Adam Smith**
*The University of North Carolina at Chapel Hill*

A fundamental principle within human memory research is the idea that repetition (i.e. multiple presentations of a stimulus) and testing (i.e. preliminary recall tasks) both improve recall performance. However, recent evidence suggests that in certain conditions repetition and testing can actually *decrease* item recall (Peterson, 2011). This study sought to determine whether these negative effects of repetition and testing would be more appropriately accounted for in the context of an encoding explanation or a retrieval explanation – in other words, whether the cause for decreased performance was related to how efficiently items were encoded or how effectively relevant self-cues were used during the recall task. Two experiments were designed to test these explanations by using lists of rhyming cue-target word pairs (e.g. "Beg – Leg") as stimuli. The target words of these pairs were organized pseudo-randomly in some phases and categorically in others. The ordering of these phases was intended to direct what relational information would be most salient – with initial pseudo-randomized ordering, within-pair (rhyming) similarities should be more apparent, and with initial categorical ordering, between-pair (categorical) similarities should be more easily noticed. Results of the experiments support an encoding account for the negative repetition effect, but a retrieval explanation for the negative testing effect.

A well-known proverb concerning the achievement of excellence is "practice makes perfect." Though few would truly assert that literal perfection can be attained through repetition, it is commonly accepted that repetition boosts one's performance on a wide variety of tasks. Johnstone, Ashbaugh, and Warfield (2002) were even able to demonstrate improvement in the development of writing skills – a complex cognitive task – merely through repeated practice. This use of repeated practice has been particularly prevalent in pedagogical settings, wherein the drill-and-practice technique (going over information until it is mastered) is a commonplace procedure in the teaching of academic skills – particularly those of mathematics and grammar. In a study by Brophy (1986), this strategy was especially effective among students with lower academic achievement, which lends some merit to the frequent use of the repetition-based technique. Furthermore, Brophy also observed that the usefulness of this approach is not confined exclusively to the instruction of basic skills, suggesting instead that a structured environment can be applied to "…any body of knowledge or set of skills that has been sufficiently well organized" (p. 1076). Though the idea behind the benefits of repetition is an intuitive one, this does not mean that it is void of scientific grounding. Simply put, the repetition effect merely posits that increased frequency and exposure to stimuli increases the later recall of said stimuli.

Indeed, the assertion that repetition impacts recall appears so self-evident that it hardly seems worthy of mention. However, it is worth noting because this principle alone cannot account for differences in individuals' levels of recall in practical settings. For instance, two students may spend the same amount of time studying in preparation for a test, but this in no way ensures identical performance on the exam. One must also take into account the various mnemonic strategies that can be implemented to increase the effectiveness of repetition as a

learning tool. For instance, it has been observed that information is better retained if practice is temporally distributed (i.e. "spaced rehearsal") than if it is presented with higher frequency during a shorter time interval (i.e. "massed rehearsal). In a study by Dempster (1987), researchers found that the benefits of this spacing effect apply to subjects who are actively attempting to learn new and unfamiliar vocabulary. This finding is particularly compelling because it displays the utility of the spacing effect not only for memorizing familiar stimuli, but also for learning new material, thereby bolstering the assertion of its usefulness as a pedagogical tool.

In more recent years, researchers have been looking into a somewhat less explored phenomenon known as the "testing effect." Although tests are typically used to gauge the retention of knowledge, there is strong evidence to suggest that testing itself actually alters memory traces and affects later recall. More precisely, the testing effect refers to the observation that testing augments a participant's retention of information more effectively than simply restudying the material. In a review of the testing effect, Roediger and Karpicke (2006) further subcategorize this phenomenon into two forms: a mediated testing effect and a direct testing effect. Mediated testing effects are those which indicate that, "it is not the act of taking the test itself that influences learning, but rather the fact that testing promotes learning via some other process or processes." (Roediger & Karpicke, 2006, p. 182). In contrast, a direct effect of testing would attribute the increased retention of information with the act of taking a test itself rather than some alternative mediating process.

The general result of experiments which explore this effect is that a group given an initial pretest outperforms a control group (with no initial pretesting) on a final measurement of information recalled. This effect persists even in instances where there is no feedback provided after initial testing, decreasing the likelihood that this recorded improvement is actually due to some mediating factor caused by such feedback. For the purposes of the current study, reference to the testing effect will specifically indicate this direct form of the effect.

There is another effect that is similar in nature (and results) to the testing effect known as the generation effect. The generation effect refers to the observation that generating information from past knowledge typically results in greater memory retention than simply reading the same material. In an experiment designed to test this effect, subjects are prompted to generate word of interest when given a meaningful cue. For instance, if the cue was based on antonyms, participants might be presented with the stimulus *"hot-c___"*, and be expected to generate the word *"cold"* as a response to this cue (Mulligan & Lozito, 2004). Typically, a generation condition yields better performance on subsequent memory tests as compared to a control condition in which participants are merely instructed to read the word pairs. Although this is comparable to the testing effect, it is important to distinguish these two phenomena. Chiefly, it should be noted that the testing effect occurs as a result of a participant accessing his or her episodic memory in order to bring to mind an item which was depicted during prior study. In contrast, the generation effect is prompted when a target item is being generated from semantic memory in relation to a given cue (in other words, not retrieved from a specific study phase, but rather generated based on general knowledge).

*The negative repetition effect and negative testing effect* The positive effects of utilizing both repetition and testing to improve memory are quite well documented. However, it is important to determine if these effects are always positive. For instance, might it be possible that repetition of verbal stimuli could produce a *negative* effect on memory in certain conditions? At first glance this proposition seems not only counterintuitive but also unlikely, as it appears incongruent with the majority of memory research that has been conducted thus far. Nevertheless, despite the apparent consensus concerning positive effects of repetition on memory storage and retrieval, evidence that may call some aspects of the well-established repetition effect into question has recently surfaced.

In Daniel Peterson's (2011) recent dissertation, a number of experiments were conducted in order to determine how the "item-specific" versus "relational" account could explain certain elements of the testing effect. In short, this account holds that informational qualities of a set of stimuli are processed in two basic ways: by attending to features which are unique to a particular stimulus (i.e. item-specific processing), or by attending to features which are commonly expressed by a set of stimuli (i.e. relational processing) (Hunt & McDaniel, 1993). In interpreting the data from these experiments,

Peterson discovered an unpredicted result. Namely, participants in one condition were presented with a list of words twice, and yet they recalled 13% fewer target words than subjects in another condition who were given the same list only once. This finding is particularly surprising considering that the group given the word list twice experienced each stimulus in a spaced manner – a presentation that should have maximized the advantages of repetition. Peterson referred to this result as a "negative repetition effect".

Considering this finding, it seems reasonable as well to question whether the testing effect is always positive. Indeed, in this same dissertation, Peterson (2011) offers evidence that contradicts this invariably positive outlook on the testing effect. However, unlike the negative repetition effect, evidence for the negative testing effect was a predicted outcome of the study. The reason for this conjecture is Peterson's hypothesis that the item-specific versus relational account, which has been used to explain the similar generation effect, can explain the testing effect. Peterson surmises that because the generation effect can be shown to negatively affect memory under certain conditions, if a similar experiment were to be conducted with the testing effect replacing the generation effect, the observation of a negative testing effect would strengthen the idea that both of these phenomena can be explained via the item-specific versus relational account. In an effort to test this prediction, Peterson modeled Experiment 4 of his study after a design implemented by Burns (1990). In the study, the stimuli consisted of 36 rhyming cue-target word pairs (e.g. "Beg – Leg"). Notably, the target words of these pairs fell into one of six distinct categories; using the above example, "Parts of the Human Body" would be the target word category.

This experiment consisted of two conditions and three phases. The first condition was a "restudy" condition, wherein the participant was instructed to read through a list of word pairs two times. The second condition was a "retrieval" condition, in which the participant read through the word pair list once for the first phase, but was later asked to recall the target word of each individual pair when presented with the cue word. By comparing the performance of the retrieval condition with that of the restudy condition, Peterson would be able to determine what effect testing had on the recall of target words, whether positive, negative, or null.

In phase one of the study, subjects in both conditions were presented with the cue-target word pairs in pseudo-random order, so that no two pairs with target words from the same category appeared sequentially. In phase two, the word pairs were organized by category so that pairs with target words from the same category occurred in sequence. For the restudy condition, both words in the pair were simply presented an additional time (albeit in categorical order), and for the retrieval condition the cue word was presented in order to prompt recall of the corresponding target word. Phase three was a free recall test of the target words from the earlier pairs and was the same for both conditions. As predicted, the implementation of randomized word pairs before phase two caused the retrieval condition to perform more poorly than the restudy condition, indicating a negative testing effect.

Peterson believed that the observed negative testing effect could be explained when put in the context of the item-specific versus relational account. He asserted that when stimuli are deemed unusual (in this instance, an incomplete word pair that requires retrieval of a target word), more attention is allocated to each stimulus, thereby prompting greater item-specific processing. Additionally, this allocated attention also results in increased processing of the relationship between the cue and target of a given word pair. Unfortunately, due to the limited nature of cognitive resources, this heightened level of individual processing inhibits relational processing between target words from different pairs. Due to this inhibition, the relational similarities between targets are far less salient in this condition. As a result, target words were more difficult to recall during testing in this condition due to the overemphasis of encoding for item-specific information as opposed to relational information between word pairs.

Although the negative testing effect was a predicted outcome of Experiment 4, comparing these results with an earlier experiment in Peterson's (2011) study yielded an unexpected result. In Experiment 3, the negative generation effect was being examined. The control condition of this experiment was a single-presentation condition, wherein subjects read a categorically organized list of rhyming cue-target word pairs once before recall testing. The unpredicted finding was that this single-presentation condition produced higher performance results than did the restudy

condition of Experiment 4, indicating the presence of a negative repetition effect.

*Possible explanations* At this point we will examine the two prominent explanations for these aforementioned negative effects on memory. The first of these is the notion that the negative testing and negative repetition effects are a result of how the presented information is encoded by the participant. We will begin with the negative repetition effect. In Experiment 4 of Peterson's dissertation, the restudy group was presented with the same list of rhyming word pairs twice. During the first presentation, these word pairs were presented randomly; the following presentation sorted the word pairs into categories based upon the target word. The single-presentation group in the accompanying Experiment 3 (analyzed post-hoc as a control) that was compared to this condition was presented with the word pairs in categorical groups, but did not have a prior random presentation of the items.

In addition to the item-specific versus relational account described earlier, another explanation for the reason the Experiment 4 restudy group did worse is the principle of negative transfer. Negative transfer is the concept that ineffective encoding strategies may be transferred from one list presentation to another, thereby reducing memory performance. To provide a comparison, the occurrence of negative transfer in memory encoding situations is analogous to functional fixedness in creative problem-solving (wherein suboptimal problem-solving strategies persist from one situation to the next).

To apply this concept as an explanation, one might reasonably propose that the cause for poorer recall in the restudy group is the greater amount of attentiveness subjects gave to the within-pair similarity (namely, the fact that the word pairs rhymed). Due to the fact that the first presentation was pseudo-randomized, it is unlikely that participants would have noticed the relational properties between the target words because there was no organized grouping of items. Because this between-pair categorization was unlikely to be noticed initially, the subjects might have been biased to notice only the within-pair rhyming similarity in the subsequent presentation. In contrast, the group that was only presented the items once (but in categorical groupings) might have been more likely to notice the between-pair relational information. By encoding the target words within the context of meaningful categories, it is reasonable to assume that these subjects would gain an advantage during the free recall phase. Although repetition and the spacing effect would normally create higher recall of targets at test, such benefits were not enough to surmount the deficit caused by encoding the target items as unrelated.

There is, however, the possibility that encoding is not solely responsible for these observed discrepancies between groups. It may be the case that the observed negative repetition and negative testing effects are retrieval-based phenomena. Simply put, this notion would assert that during the final free recall test, participants in the "restudy" (repetition) and "retrieval" (testing) conditions of Experiment 4 were less able to actively recall the categories observed earlier. Due to this deficiency, they would be unable to cue themselves to enhance recall ability. However, the condition in the corresponding Experiment 3 that was only presented with the word pair stimuli once and in categorical order (the single-presentation condition) would have a higher likelihood of retrieving the category information upon testing, and as such would be able to provide self-cues to reduce the difficulty of recalling the targets during the final recall task. This alternative was not explored in Peterson's dissertation, but it remains a possible explanation for what may have been prompting these observed negative effects.

*The current study* Although Peterson's observations offer a compelling and unique perspective on the nature of repetition and testing effects, they are not conclusive. Peterson (2011) does not deny this, and provides a detailed account of limitations that are apparent in his study. First among these is the fact that the explanation for the negative repetition effect is entirely post-hoc. Having not set out to test this phenomenon – indeed, he was surprised himself that it occurred – he simply analyzed the data after the unexpected trend was noted. Furthermore, he again notes that the analysis involved a cross-experimental comparison. Although the main effect of the two experiments was found to be significant and the populations utilized were quite similar, there was no random assignment of subjects to conditions. Peterson believes that accounting for these two points would make this finding far more compelling (*see* Peterson & Mulligan, in-press).

With respect to the negative testing effect, Peterson's findings were more convincing, as a majority of his experiments were successful in producing the effect. Nevertheless, this study is only scratching the surface of the potential implications a negative testing effect might represent. Although there is a considerable amount of literature on the testing effect, Peterson laments that precious little is known about the process itself. As such, he feels that further emphasis should be placed on determining the mechanisms underlying the testing effect to determine exactly *why* it is that tests facilitate improvement in (or in this case, inhibition of) memory.

The current study was designed as a means of addressing these issues through the course of two experiments. In addition, the design of both experiments is also intended to determine the source of the negative testing and negative repetition effects – in other words, whether an encoding or retrieval explanation can more effectively account for the occurrence of either effect. In order to achieve this, the structure of both experiments will be based upon Peterson's (2011) dissertation as well as previous studies (Burns, 1990; Karpicke & Zaromb, 2010; Roediger & Karpicke, 2006).

For Experiment One, there will be a single-presentation control condition which will require the subject to read over a list of rhyming word pairs once. These word pairs will be organized by the category of the target word; for instance, "Linger – Finger" and "Harm – Arm" would be positioned next to one another since the target words fall within the same taxonomic category. A second condition (the "restudy" condition) will be allocated to observe how the effects of repetition compare to the control group. This condition will be presented with the same list of word pairs through two phases – first in pseudo-random ordering and then in categorical ordering. The final condition (the "retrieval" condition) will gauge the results of the direct testing effect in comparison to the restudy group. The first phase here will be identical to the restudy condition's first phase, but during the second phase participants will be asked to generate the appropriate target word when given only the accompanying cue word. For instance, if the word pair "beg – leg" was in the first phase, participants would be presented with the stimulus "beg – ___" and asked to fill the blank accordingly (importantly, the retrieval group will be given the correct answer at the end of each stimulus presentation).

Finally, Experiment One will also feature a category-cued recall test as the final phase for all conditions (as opposed to the original free recall test). By cuing all groups identically with the categories in which the earlier target words belong, the availability of the category names as retrieval cues will be equalized for all conditions. After testing, the performance of each condition will be assessed and compared. Specifically, the single-presentation and restudy conditions will be compared in order to assess the effect of repetition, and the restudy and retrieval conditions will be compared to determine the effect of testing.

The reason the comparisons are made in this way (as opposed to having both the restudy and retrieval conditions compared with the single-presentation condition) is to isolate the structural elements that vary between each condition. In other words, there are so many similarities between the conditions that it is necessary to determine which alteration is responsible for the observed outcome. By comparing the single-presentation and restudy conditions, the only difference between the groups is the inclusion of a pseudo-random word pair presentation in the latter. By comparing the restudy and retrieval conditions, the only difference is how the organized list of phase two is presented (i.e. keeping the target words initially blank so the participant has to recall the word from memory). If the retrieval condition was compared to the single-presentation condition, it would be unclear whether the results were due to the element of repetition (also found in the restudy group) or the testing element of phase two. Rather, the retrieval condition will be compared with the restudy group to determine whether the inclusion of a testing element compounds the negative effect with the repetition already inherent in both conditions.

If either (or both) of the negative effects are still present, this will provide evidence against the reliance of retrieval cues in mediating the effect(s). However, if a given negative effect does not appear after the category-cued recall test, this would suggest that retrieval plays a larger role in the facilitation of the given effect than was originally presumed.

The design of Experiment Two will be similar in nature to the first experiment with a few notable changes. First, the restudy condition will feature the categorical ordering of cue-target word pairs *before* the pseudo-randomized presentation, essentially interchanging the first

two phases of the Experiment One restudy group. Since the same phase substitution cannot be made for the retrieval condition – wherein phase two relies upon prior study of word pairs – this group will be omitted from the experiment. Finally, the category-cued recall test at the end of each condition will be replaced with a free recall test.

Comparison of these two groups will once again indicate the effect of repetition under these experimental parameters. However, in this instance the compared results have different implications. For instance, if the negative repetition effect is not present and the restudy group has comparable (or perhaps superior) performance to the single-presentation condition, this would indicate a greater likelihood of an encoding phenomenon taking place. This explanation is derived from the fact that presenting the categorically organized list of word pairs in phase one of the restudy condition should mitigate or eliminate the occurrence of negative transfer, thereby allowing for more efficient encoding of relational information in phase two (and consequently higher performance on the recall test). In contrast, if the negative repetition effect does indeed occur in this experiment, this finding would suggest that an encoding explanation is less capable of accounting for the results, suggesting the possibility of a retrieval-based explanation.

## EXPERIMENT ONE

### METHOD

*Participants* Sixty-eight participants were obtained through the Introductory Psychology subject pool at the University of North Carolina at Chapel Hill. Time spent during this experiment was allocated to each participant as a number of laboratory credits necessary for their class. There were 23 participants in the "single-presentation" condition, 23 participants in the "restudy" condition, and 22 participants in the "retrieval" condition.

*Materials* The critical items were a set of 36 rhyming cue-target word pairs borrowed from Peterson's (2011) dissertation (p. 56). The target words of these pairs fell into one of six different taxonomic categories, each containing six exemplars of the given category. These targets were borrowed from the category norms of Van Overschelde, Rawson, and Dunlosky (2004) – an updated and expanded list of category norms originally assembled by Battig and Montague (1969). The six categories were: "Parts of the Human Body", "Vehicles", "Kitchen Utensils", "Fruits", "Animals", and "Metals". This is also the order in which they were presented during the categorical presentation for all groups – phase one for the single-presentation condition, phase two for the restudy and retrieval conditions. To complete the word pairs, a rhyming cue-word was assigned in conjunction with each of the target words; however, the cue-words were not themselves a member of any of the six target categories (to avoid potential intrusion caused by accidental cue-word recall).

Once these word pairs were assembled, they were organized into two different lists. The first list was organized so that the target words were presented sequentially in a pseudo-random series. The purpose of pseudo-randomization of the word pairs (as opposed to unrestricted randomization) was to ensure that no two pairs with target words from the same category appeared in succession. This pseudo-randomized list appeared in phase one of the restudy and retrieval conditions, but did not appear in the single-presentation condition. The same word pairs were then assembled into a categorically organized version of the list, wherein the target words were grouped serially in relation to their taxonomic category. This list was introduced in phase one of the single-presentation condition and phase two of the restudy and retrieval condition. There was one notable distinction for the organized list in the retrieval condition; specifically, participants in this group were initially provided only the cue-word and a blank for the target word. The purpose of this was to allow subjects a period of time where they would attempt to retrieve the corresponding target from the earlier (pseudo-randomized) presentation of the word pairs.

*Procedure* Upon arrival, subjects were informed of their rights as research participants and then asked to sign two copies of the IRB consent form. Experimental sessions occurred with one participant and one experimenter per session (i.e. multiple trials were not conducted simultaneously).

Participants in the restudy condition underwent three experimental phases. In phase one, the participants were presented the pseudo-randomized list of rhyming word pairs

(described above). The experimenter briefly outlined the "cue-target" nature of the word pairs, and participants were instructed to read the pairs silently. They were also informed that they should attempt to learn the rhyming pairs of words for a later memory test. The word pairs were presented in the center of a computer screen in black lettering on a solid white background. Each word pair was presented individually for four seconds followed by a 500 millisecond interstimulus interval taking the form of a solid white screen. After the first phase was completed, the participant was given a math distractor task consisting of 70 arithmetic problems (which did not deviate from standard four-function mathematical notation). The participant was allotted five minutes to complete as many mathematical problems as possible without making any notes or intermediate calculations.

After the time for the distractor task elapsed, phase two of the restudy condition began. In this phase, the categorically organized version of the word pair list was utilized. Participants were reminded that later in the experiment they would be asked to remember information presented, but in this instance were specifically asked to remember the target words. Word pairs were presented more slowly in this phase (15 seconds each), and participants were asked to read each pair aloud. After this list was completed, phase three began. In this final phase, the participant was given a category-cued recall test. The participant was told that the target words came from different categories, and that category names would be presented to help the participant recall the target words. Each categorical cue was presented on the computer screen for 50 seconds for a total testing duration of 5 minutes. Participants were asked to recall as many target words that came from the category as they were able. It was made clear that each category corresponded to multiple target words, and that participants should try to recall as many targets as possible for each category. After this task was completed, the sheet with the recalled target words was collected, the participant was debriefed, and the appropriate amount of laboratory credit was assigned.

The next condition to outline is the retrieval condition, which differed in only one aspect from the restudy condition. During phase two, participants in the retrieval condition were presented with the same categorically ordered word pair list described above, but with the target word missing (i.e. the cue-word was presented in isolation). Participants were instructed to read the cue word aloud, and then say the name of the target word aloud once they recalled it. After 10 seconds, the target word was presented. If the participant incorrectly recalled the target or did not recall any target word, s/he was asked to read the target aloud. The full word pair then remained on the screen for five seconds. This was followed by a 500 millisecond interstimulus interval. While the participant was reading these word pairs aloud, the researcher was scoring the responses. A correct vocalization of the target word before it was presented was coded as a "correct response." An incorrect vocalization of the target word was coded as an "incorrect response" (the experimenter recorded the incorrect word for this response). Failure to vocalize either an incorrect or correct target word before it was presented was coded as "no response." In the rare instances in which a participant vocalized an incorrect target word followed by the correct target word, both the incorrect and correct responses were recorded (with the understanding that the participant was initially incorrect but provided a revised answer before the target word was presented). All other aspects of this condition were identical to the restudy condition.

The final condition of this experiment was the single-presentation condition. This condition was identical to the restudy condition except for the omission of phase one (and the corresponding distractor task). In other words, the only word pair list studied was the categorically organized version, and it was only studied once before the category-cued recall test. All other aspects of this condition matched the design of the restudy group.

## RESULTS

In order to compare the proficiency on the category cued recall test across the three conditions, a one-way between subjects ANOVA was conducted. For all statistical tests of significance, an alpha level of .05 was utilized. The results of the analysis indicated a significant difference in the number of items correctly recalled between the single-presentation ($M$ = 25.00, $SD$ = 5.71), restudy ($M$ = 20.44, $SD$ = 7.83), and retrieval ($M$ = 19.36, $SD$ = 4.28) conditions; $F(2, 65) = 5.398$, $p = .007$ (Figure 1). Post hoc comparisons between individual

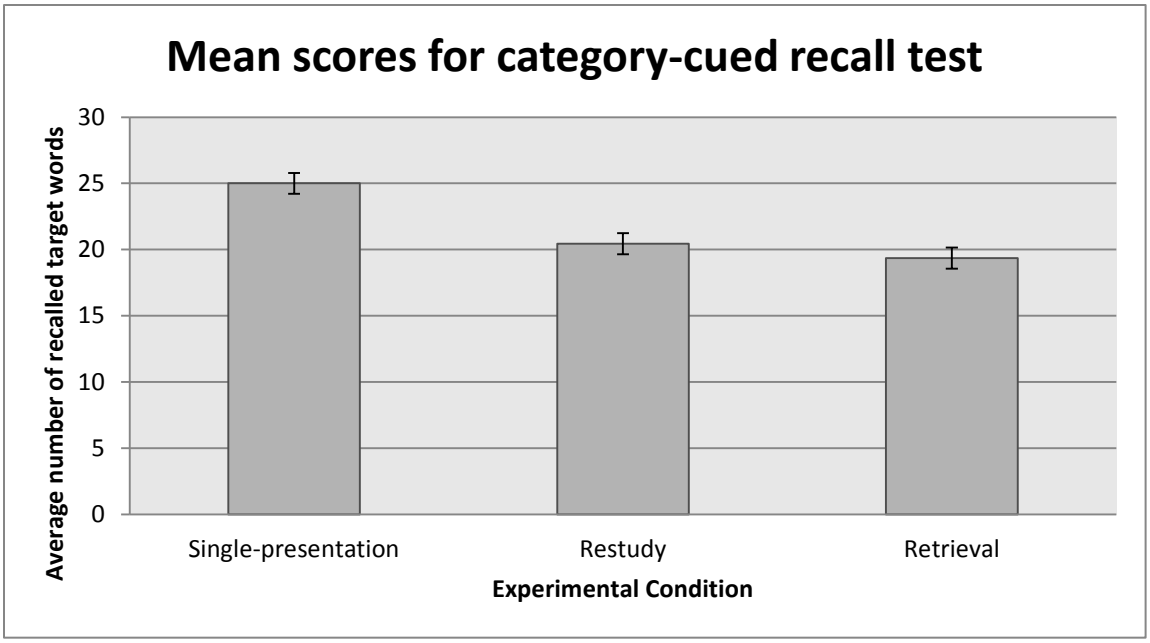


FIGURE 1. The compared performance on the category-cued recall test of Experiment One. There was a significant difference between the single-presentation and restudy conditions (p = .014), but not between the restudy and retrieval conditions (p = .561). Error bars represent average standard error of the means (± 0.792).

groups were assessed using Fisher's LSD, and indicated a significant difference between the single-presentation and restudy conditions (p = .014). However, the difference between the restudy and retrieval conditions was not found to be significant (p = .561). In other words, the single-presentation group performed significantly better than the restudy condition (suggesting the presence of a negative repetition effect), but the restudy condition failed to garner significantly higher scores than the retrieval condition (suggesting an absence of a testing effect, either negative or positive).

We also needed to determine if participants in the conditions were using different response criteria (e.g. engaging in varying levels of guessing). In order to accomplish this, a one-way ANOVA was conducted to compare the frequency of intrusions between the three conditions. Intrusions were classified as items that were reported by participants during the category cued recall test that were not actually presented during the study phase(s), but were associated with a given category. Results of the ANOVA indicated that there was not a significant difference in the mean score of intrusions among the three groups; $F(2, 65) = 1.117$, $p = .333$. Due to this observation, we can safely claim that the occurrence of intrusions did not impact the performance of any one condition significantly more than another.

## DISCUSSION

The outcome of Experiment One revealed the occurrence of one negative effect but the absence of the other. Namely, while the negative repetition effect was found to occur under these experimental conditions, the negative testing effect was not. Although both the restudy and retrieval conditions performed significantly worse than the single-presentation condition, this comparison was only relevant for the restudy group. In order to provide evidence for a negative testing effect, the retrieval condition would have to have performed significantly worse than the restudy group (since both of these conditions contain a form of repetition, this comparison isolates the testing group's definitive methodological variation in phase two).

These results support the use of an encoding account to explain the observed negative repetition effect. The encoding explanation predicted that the introduction of category cues in the recall test would not prevent a negative effect from occurring. This

prediction was based upon the assumption that some conditions will facilitate greater encoding of between-pair relational information (i.e. categorical processing) than others during the study phase(s), and that this encoding variation is what accounts for the observed difference in scores. Since the presence of category cues did not eliminate the negative repetition effect, it is reasonable to infer that the difference between the single-presentation and restudy conditions arose instead during the encoding phase(s) – in other words, there was differential encoding of between-pair relational information.

In contrast, the failure to produce a negative testing effect in this experiment can be accounted for by the retrieval explanation. The retrieval explanation predicted that a recall test with category cues would prevent a negative memory effect from occurring. When participants in the restudy and retrieval conditions were supplied with the same category cues, the retrieval condition was no longer at a comparative disadvantage. This finding implies that the categorical similarities between-pairs were encoded equally in these two groups. Therefore, previous variation of performance between these two conditions – such as in Peterson's (2011) Experiment 4 – should be attributed to the reduced use of category information during recall for the retrieval condition.

## EXPERIMENT TWO

### METHOD

*Participants* Fifty-two participants were obtained through the Introductory Psychology subject pool at the University of North Carolina at Chapel Hill. Time spent during this experiment was allocated to each participant as a number of laboratory credits necessary for their class. There were 26 participants in the "single-presentation" condition, and 26 participants in the "restudy" condition.

*Materials* For consistency, the 36 rhyming cue-target word pairs used in this experiment were identical to those used in Experiment One. Likewise, the pseudo-randomized and categorically grouped versions of the lists used in the previous experiment were the same ones used here.

*Procedure* The procedure of Experiment Two matched Experiment One in most respects, but varied in a few important ways. First, this experiment did not feature a retrieval condition, and as such made no assertion as to the potential causes for the negative testing effect. Second, in the restudy condition, phase one contained the categorically grouped cue-target list and phase two contained the pseudo-randomized list, reversing the previous order of grouping presentation.

Finally, instead of using a category-cued recall test, the last phase for both the single-presentation and restudy conditions took the form of a free recall test. Participants were provided a blank sheet of paper and pen with which to record the target words presented in the earlier phase(s) of the experiment. Furthermore, they were instructed to record target words in the order that they were recalled. The testing period lasted 5 minutes.

*Results* For the second experiment, an independent samples t-test was used to compare the performance on the free recall test in the single-presentation and restudy conditions. Again, all statistical tests of significance used a .05 alpha level. In this instance, there was not a significant difference found in the scores for the single-presentation ($M$ = 21.00, $SD$ = 7.56) and restudy ($M$ = 23.62, $SD$ = 8.59) conditions; t(50) = -1.166, p = .249 (Figure 2). This finding indicates that switching the order of study phases in the restudy condition results in performance which does not significantly differ from the control (single-presentation) condition; in short, a negative repetition effect was not present in this experiment. On the contrary – the extent of any trending identified in the recall data is actually in the direction of a *positive* repetition effect.

Again we needed to determine if participants in the conditions were using different response criterions. Given that this experiment utilized a free recall test, there were two primary recall deviations considered. One of these was a participant's accidental recall of a cue word instead of a target word. A t-test revealed that there was not a significant difference in the occurrence of this mistake between the single-presentation ($M$ = .269, $SD$ = .667) and restudy ($M$ = .50, $SD$ = 1.14) groups; t(50) = -.891, p = .377. Intrusions were also recorded in this experiment, and were found to be equivalent for the single-presentation ($M$ = .462, $SD$ = 1.029)

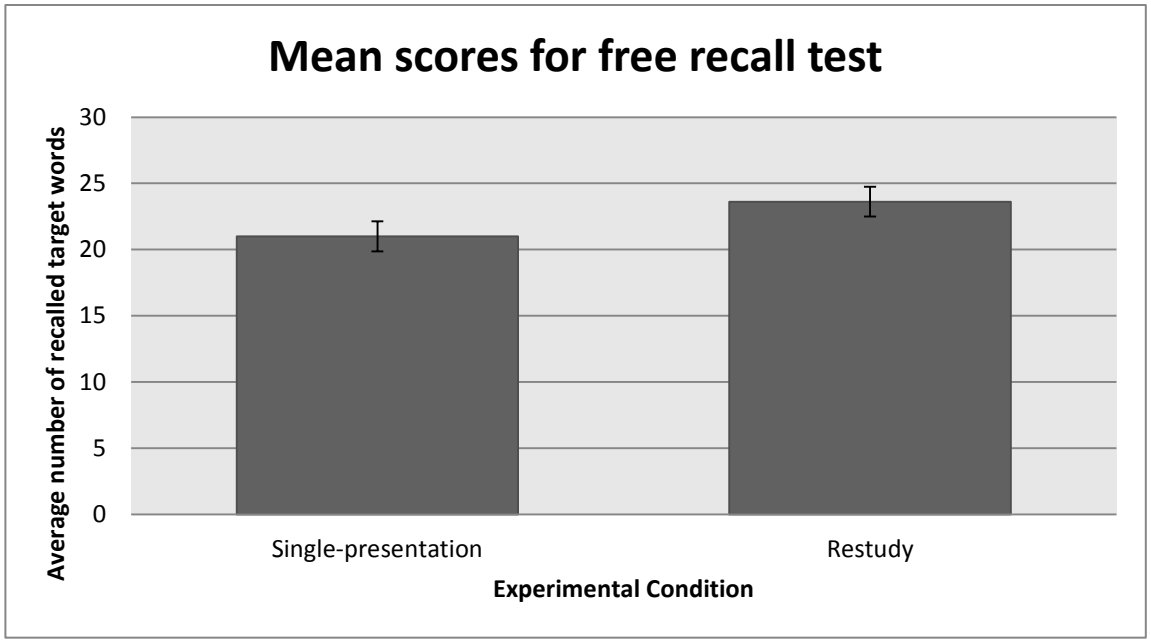


FIGURE 2. The compared performance on the free recall test of Experiment Two. The difference in scores between these groups was non-significant ($p = .249$). Error bars represent average standard error of the means (± 1.126).

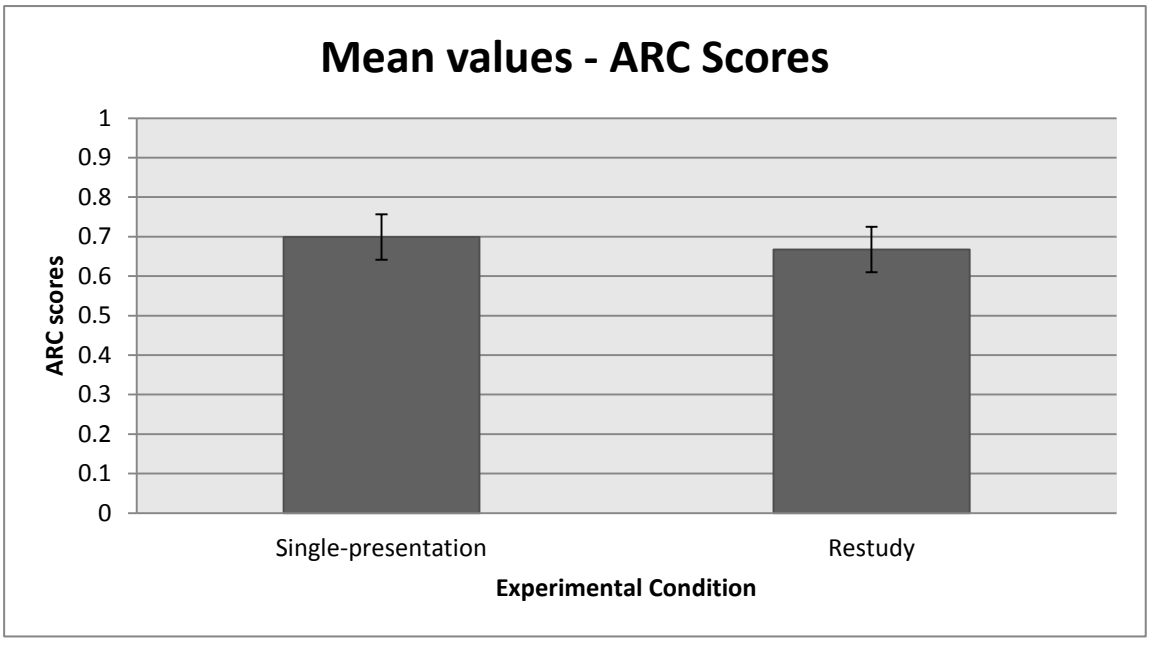


FIGURE 3. The compared adjusted ratio-of-clustering (ARC) scores between the two conditions. The difference in scores between these groups was non-significant ($p = .788$). Error bars represent average standard error of the means (± 0.0578).

and restudy ($M$ = .192, $SD$ = .402) conditions; t(50) = 1.243, p = .220.

Finally, adjusted ratio-of-clustering (ARC) scores were computed by assessing the frequency with which target words of the same category were recalled in succession during the free recall test. This metric is of particular interest as it indicates either the presence or lack of categorical grouping of target items (beyond levels of chance) by subjects in a given condition. Specifically, an ARC score of 0 indicates chance-level clustering of target items, positive scores indicate above-chance frequency of clustering, and a score of 1 means that all target items were perfectly clustered (i.e. all targets were grouped in categorical succession). An independent samples t-test indicated that there was not a significant difference in ARC scores between the single-presentation ($M$ = .70, $SD$ = .4326) and restudy ($M$ = .67, $SD$ = .4084) groups; t(50) = .270, p = .788 (Figure 3). However, the scores for both conditions were positive, indicating a frequency of item clustering in both groups that surpasses levels of chance. Therefore, we can infer that while participants in the two conditions did not significantly differ in the practice of categorically grouping the target items during free recall, both conditions made use of category clustering during the recall test.

## DISCUSSION

The results of Experiment Two indicated the absence of a negative repetition effect in these experimental conditions. In other words, the restudy condition did not perform worse than the single-presentation condition. On the contrary, the restudy group actually garnered a higher average score (although this difference was not statistically significant). These results are consistent with an encoding explanation of the negative repetition effect. By providing both conditions of this experiment with identical and categorically grouped initial series of word pairs, it was hypothesized that negative transfer would not occur in the restudy group (as the initial encoding was the same for both conditions). In other words, this design increased the likelihood that subjects in the restudy group would initially recognize the relational organization among the target words. Since the pseudo-randomized list was provided *after* these initial associations were made, they did not inhibit the participant's ability to group target items categorically for more efficiency in the following recall task. The comparative performance of these two groups suggests that when initial encoding is standardized for the single-presentation and restudy conditions, the negative repetition effect dissipates.

Further evidence for the similarity of encoding between the conditions comes from a comparison of the ARC scores. Since ARC scores function as an indicator of organizational processing of items in a list, the finding that the two conditions did not significantly differ on this metric suggests that both groups utilized comparable levels of categorical processing during recall. Furthermore, the results indicated that the level of target-item clustering exceeded levels of chance for both groups. In other words, not only did both groups exhibit similar levels of categorical processing, but the grouping was too organized to be accounted for by chance. Altogether, these results reflect the reliance of participants in both conditions on between-pair relational processing.

## GENERAL DISCUSSION

Taken together, these experiments successfully expanded the scope of Peterson's (2011) original study. An instance in which the negative testing effect does not occur may seem damaging at first glance, but it actually leads to a greater understanding of the conditions under which the effect surfaces by eliminating conditions under which the effect is absent. Though this result indicates a notable instance where the effect is missing, it will be the responsibility of future experiments to continue isolating instances in which the effect does take place. Furthermore, this study explored the negative repetition effect by implementing an experimental design structure, which was intended to detect such an occurrence from the start (instead of invoking cross-experimental comparisons as was done originally). The successful replication of the negative repetition effect reduces the likelihood that the original observation of its presence was anomalous, lending credence to its existence. Finally, these experiments succeeded in indicating an appropriate source for both of the negative effects – namely, the negative repetition effect was linked with the encoding explanation, and the negative testing effect was explained by the retrieval account.

*Limitations* Although these results broaden earlier claims concerning the negative effects of

repetition and testing, they are not without their own limitations which merit further exploration themselves. One factor which was not considered is how longer intervals of delay between word pair studying and testing may alter the expression of the aforementioned negative effects. In other words, would the same trends occur if the delay interval were a day (as opposed to the five minute distractor task utilized in this study)? Prior memory research suggests that by increasing the interval there should be a more noticeable distinction between groups utilizing effective study techniques (e.g. repeated testing) and those who are not (Roediger & Karpicke, 2006). However, would this same trend hold when experimental conditions are designed to cause one of these negative effects on memory? Answering this question is an important step to determining the relevance of these effects in an applied setting – if differences in performance do not persist over a longer period of time, then it may not be worth altering pedagogical strategies to account for this short-lived deficiency.

At this early stage in the research of these negative effects, another notable limitation is the lack of complexity in the stimuli being studied. By only using rhyming cue-target word pairs, the information being studied presumably forms fewer intricate associations between items than more complex stimuli might. In other words, a participant's processing of the association between two target words is notably simpler than the association between two *concepts*. Considering that much of what is studied in academia is frequently grouped into such abstractions, it will be important to determine whether or not the negative effects of repetition and testing directly influence only individual items of information, or larger bodies of knowledge as well.

Finally, as with most lines of research, this study may have benefited from a greater variety of represented demographics. Of these demographics, education and age seem to be the most immediately pertinent to account for. The entire sample of participants was comprised of undergraduates, and it is reasonable to suspect that college students may have more firmly grounded study habits than the general population. Due to this, the process of encoding the presented study items may vary in some key way, resulting in an altered expression of these effects – perhaps enhanced, perhaps reduced.

A similar issue may arise with children, who do not have as much practice with studying lists of information. It may be the case that the usefulness of relational associations for categorically similar information may not be as obvious for children, and therefore may not affect them in the same manner. If, however, these demographic groups were to behave in a manner similar to our undergraduate sample, then there would be greater support for the validity of applying this information both for early education and outside the realm of academia proper.

*Future research* One of the chief concerns for future lines of research is the construction of experiments which increasingly feature types of information which are more likely to be sequentially processed in a given setting. In doing this, the results are less confined to the purely theoretical, and modes of application become more readily apparent. Continuing with the example of educational relevance, future studies might feature a comparison between how negative effects of repetition and testing affect varying disciplines of study. For instance, it may be possible that studying vocabulary, historical facts (such as dates or event locations), and mathematical equations all prompt negative effects in study conditions similar to what was tested above. However, it may be just as likely that the nature of the material being studied moderates the potency or expression of a given negative effect. Not only would this knowledge help to shape classrooms for optimal learning, but it may also shed light on the underlying mechanisms responsible for the occurrence of the effects in the first place.

While exploring the qualitative features of the information studied in such experiments would be a meaningful step, it is also important to remember that the results of this study have indicated that the nature of the final test helps to determine whether the negative testing effect will occur. Specifically, with fewer cues the negative testing effect is found (*see* Peterson, 2011), but with useful cues explicitly provided (e.g. presentation of categories for target words) the testing effect – either positive or negative – does not occur. Considering this, future studies should utilize a variety of final memory tests in conjunction with the use of alternative forms of informational stimuli; this way, subsequent research will be able to account for experimental designs in which the negative testing effect is either expected to occur or be absent.

Studies such as these would assist in identifying how modifying qualitative aspects of

experimental design might cause either of the negative effects on memory. However, experiments which address the typical *quantity* of information retention necessary to perform well on a test would also be of value. So far the negative repetition effect and negative testing effect have exclusively been demonstrated in instances where there are only a few memorized words jotted down over the course of five minutes – in an actual testing session there tends to be greater amount and variety of information, as well as considerably more time to complete the assignment. Assuming the answers on a test could be objectively assessed as correct or incorrect (i.e. no essay or opinion based questions), it would be possible to utilize such an exam as the final recall task of an experiment somewhat resembling those conducted in the current study. For instance, historical information might be presented in such a way that emphasizes facts about individual battles rather than their context during a war (e.g. "Normandy Invasion / Eisenhower / Omaha Beach – D-Day"). Item lists such as these could be organized categorically (e.g. "World War II") or presented pseudo-randomly so that no two battles corresponding with a single war are presented in succession. When organized pseudo-randomly, the nature of these individual items should focus processing on within-stimulus characteristics which would obscure the apparent similarity between-stimuli and potentially cause negative transfer to affect subsequent presentations. Stimuli such as these are qualitatively distinct from those in earlier studies and demand a greater amount of retained information, thereby addressing both of the previously stated concerns simultaneously. Using a more natural structure such as this also helps to ensure that the negative effects of repetition and testing are less likely to be laboratory effects and will genuinely occur in more practical settings (in this instance, a history exam).

*Implications* Although the potential pedagogical implications will require further research to identify and validate, the theoretical impact of these experiments concerning the negative effects on memory recall is notably more immediate. Without speaking too broadly, studies such as this indicate that there may need to be some reassessment of the roles that repetition and testing play in memory retention. Whereas these strategies were once thought to uniformly and invariably improve – or at least not reduce – one's performance on memory tasks, it now seems that this notion is either incorrect or incomplete. The recognition and encoding of relational information which meaningfully connects stimuli in a series seems to play an integral part in the determination of whether or not one's memory will improve or diminish. In short, it is not only the frequency of exposure to information or general efficacy of studying methods which affect memory performance, but also the organizational structure of the material itself.

So it seems that in some instances practice may not always ensure perfection. Though it would be unwise to suggest abstaining from repeated practice entirely (as it is effective in far more instances than it is detrimental), it no longer seems appropriate to accept its usefulness as an undisputed facet of memory research. As future studies continue to uncover variations and exceptions to the typical patterns of memory retention, we must remember that the arrangement of information being studied plays as much of a role in memory as how we endeavor to absorb that information.

## References

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, *80*(3, Pt.2), 1-46.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, 41(10), 1069-1077.

Burns D. J. (1990). The generation effect: A test between single- and multifactor theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1060-1067.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2), 162-170.

Hunt, R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal Of Memory And Language*, *32*(4), 421-445. doi:10.1006/jmla.1993.1023

Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology*, 94(2), 305-315.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal Of Memory And*

*Language*,*62*(3),227-239. doi:10.1016/j.jml.2009.11.010

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. H. Ross, B. H. Ross (Eds.), *The psychology of learning and motivation: Advances in research and theory, Vol 45* (pp. 175-214). San Diego, CA US: Elsevier Academic Press.

Peterson, D. (2011). *The testing effect and the item specific vs. relational account* (Doctoral Dissertation). The University of North Carolina, Chapel Hill, NC.

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

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

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289-335.