January 2011

Covert Retrieval Practice Benefits Retention As Much As Overt Retrieval Practice

Megan Smith
Washington University in St. Louis

WASHINGTON UNIVERSITY

Department of Psychology

COVERT RETRIEVAL PRACTICE BENEFITS RETENTION AS MUCH AS OVERT RETRIEVAL PRACTICE

by

Megan Alice Smith

A thesis presented to the
Graduate school of Arts and Sciences
of Washington University in
partial fulfillment of the
requirement for the
degree of Master of Arts

August 2011

Saint Louis, Missouri

Acknowledgements

I wish to thank my advisor Dr. Henry L. Roediger for his invaluable guidance, and Dr. Mark A. McDaniel and Dr. Kathleen B. McDermott for serving on my committee. I also wish to thank Kelly Young for assistance with data collection and entry. This research was supported by a Collaborate Activity Grant from the James S. McDonnell Foundation, Applying Cognitive Psychology to Enhance Educational Practice: II.

Table of Contents

List of Tables

List of Figures

Abstract

A plethora of previous research shows that testing benefits retention (see Roediger & Karpicke, 2006b); however, most testing research has employed tests requiring overt responses. Does covert retrieval during testing—thinking of but not producing responses—produce the same benefit? Covert retrieval can benefit retention (e.g., Kang, 2010), but does it do so as well as overt retrieval? Tulving (1983) hypothesized that overt and covert retrieval should result in comparable retention benefits, yet research by MacLeod and collaborators (2010) suggests producing an overt response during encoding aids retention. If the same principle operates during testing, overt responding may lead to enhanced retention relative to covert retrieval. We report three experiments comparing retention after overt and covert retrieval on a first test. Experiment 1 uses a novel procedure designed to motivate subjects to retrieve during both overt and covert retrieval trials. Experiments 2 and 3 employ a procedure that more closely mirrors natural retrieval processes. The results generally confirm Tulving's hypothesis: overt and covert retrieval result in comparable retention benefits. Students can learn as much from covertly self-testing as they would from overt responding.

The Testing Effect Without Overt Retrieval

Research dating back a century has shown that taking a test is not a neutral assessment of memory. Instead testing, or retrieval practice induced via testing, is a potent way to improve memory (Abbott, 1909; see Roediger & Karpicke, 2006b for a review). Research scrutinizing the direct effects of testing on retention has found that testing improves learning and retention in a number of situations (but see Roediger, Agarwal, Kang, & Marsh, 2010 for a discussion of the boundary conditions). For example, testing has been shown to benefit retention using a range of materials, such as word lists (e.g., Carpenter & DeLosh, 2006; Karpicke & Roediger, 2007), foreign language vocabulary words (e.g., Carrier & Pashler, 1992; Karpicke & Roediger, 2008), short text materials (e.g., Roediger & Karpicke, 2006a), longer text materials (e.g., Kang, McDermott, & Roediger, 2007), pictures (e.g., Wheeler & Roediger, 1992), Chinese characters (Kang, 2010), video lectures (Butler & Roediger, 2008) and even natural categories (e.g., Jacoby, Wahlheim, & Coane, 2010). Various testing formats can improve retention, including free recall, cued recall, short-answer, and multiple-choice tests (see Carpenter & DeLosh, 2006, and Kang et al., 2007). Research has also shown that testing can benefit retention in practical settings, such as middle-school classrooms (e.g., McDaniel, Agarwal, Huesler, McDermott, & Roediger, 2011) and college courses (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). Because the direct effects of testing on retention are primarily positive and robust, and these effects have replicated in educational settings, many cognitive psychologists have recommended that testing be used as a way to promote learning in the classroom (for an example, see McDaniel,

Roediger, & McDermott, 2007). Even more recently, Roediger, Putnam, and Smith (in press) discuss ten general benefits to using tests in educational settings.

Researchers examining the retention benefits of retrieval practice have almost exclusively employed tests with overt responding. In other words, subjects nearly always take a test during which they are required to produce an overt response by writing, typing, or speaking their responses out loud. Covert retrieval—bringing information to mind or mentally rehearsing it—has scarcely been used in prior research. This is of course not without good reason. Researchers are often interested in performance on the initial tests because success on the initial tests is important for obtaining the positive effects of testing (see, for example, Butler, Marsh, Goode, & Roediger, 2006). If subjects do not produce an overt response during initial testing, then performance cannot be measured and the researcher cannot know what the subject is doing during the test. However, it is possible that covert retrieval benefits retention just as much as overt retrieval does, and producing the overt response is not necessary to obtain positive retrieval effects. In the experiments reported here I investigated whether covert retrieval produces retention benefits comparable to overt retrieval. Although a few papers have discussed the benefits of covert retrieval in the past, the literature on this topic is mixed with some finding that covert and overt retrieval produce similar retention benefits and others finding overt retrieval to be the superior method. These mixed findings may be due to a difference in the quality and amount of information retrieved during overt and covert retrieval. Relative to overt retrieval instructions, subjects may not be as motivated when instructed to covertly retrieve because they do not need to produce anything for the experimenter. Therefore, in order to examine the relative retention benefits of overt and

covert retrieval, I created a novel procedure in order to motivate subjects to retrieve during both overt and covert retrieval trials.

Whether bringing information to mind produces a retention benefit comparable to overtly producing a response is an interesting question relevant to education. Cognitive psychologists have recommended that educators encourage students to practice retrieval as a study strategy in the classroom and on their own to improve learning (e.g., McDaniel et al., 2007). If covert retrieval fails to improve retention as well as overt retrieval, then instructors should take care to implement activities requiring overt retrieval practice during learning, and students should make sure to overtly retrieve during self-testing. However, if it is the act of retrieval—bringing information to conscious awareness—that is beneficial for retention, and an overt response does not further enhance this benefit, then any activity that elicits retrieval practice, whether overt or covert, should improve retention. One criticism of using testing as a learning tool is that testing—creating the tests, administering the tests, and grading the tests—takes a lot of time (see Roediger et al., in press). If covert retrieval produces a comparable benefit, then a formal test need not be created nor graded. Instead, educators could pose questions to the class and ask all of the students to simply covertly retrieve the answer. In addition, students could employ the 3R method—a method where students first read to-be-learned material, recite or rehearse the material (i.e., practice retrieval), and then review the material again before moving on to read the next section of the material (see McDaniel, Howard, & Einstein, 2009)—using silent mental rehearsal in class or in the library on their own without taking the time to write out their responses.

Questions regarding the effectiveness of covert retrieval are still relatively open, with only a few references to its possible effectiveness sprinkled throughout the literature. Tulving's (1983) intuition was that covert retrieval works just as well to improve retention as overt retrieval. He asserted that "retrieval of information from episodic memory in response to implicit or self-generated queries—'thinking about' or reviewing the event in one's mind—produces consequences comparable to those resulting from responses to explicit questions" (p. 47). According to this hypothesis, the act of retrieval produces the mnemonic benefit on retention and the overt response does not add to this benefit. Therefore, educators need not worry about implementing retrieval tasks requiring overt responses. Tulving did not present any data to support his claim. However, a few papers examining the benefits of testing that have employed covert forms of retrieval might suggest that Tulving's assertions were correct.

For example, Orlando and Hayward (1978) examined the effectiveness of the read-recite-review (3R) method (described above) when students mentally rehearsed the information presented in a text. The 3R method is a shortened version of Robinson's (1941) survey, question, read, recite, review (SQ3R) method. This method is an effective way for students to practice retrieval during learning in order to improve retention, and is easy for students to implement themselves (see McDaniel et al., 2009). Orlando and Hayward instructed their subjects to read a 10-paragraph text and complete one of three study techniques. Some subjects used the 3R strategy: they read the text one paragraph at a time, mentally rehearsed the information, and then reviewed the paragraph before moving on to the next paragraph. Some subjects read and reread each paragraph, and others read and took notes on each paragraph. The 3R strategy with mental rehearsal

improved performance relative to rereading the material on an immediate but not a delayed test. Later performance for the group that used the 3R strategy did not differ from that of the group that took notes over the text. These results indicated that, at least on an immediate test, practicing covert retrieval during learning can improve performance relative to simply rereading the material.

Two recent papers also showed that covert retrieval benefits retention (Carpenter & Pashler, 2007; Kang, 2010). Carpenter and Pashler (2007) showed that testing could improve visuospatial map learning. In their experiment, subjects studied maps containing a number of different features. During the initial test, subjects were given an incomplete version of the map and were instructed to form a mental image of any missing features. Covert retrieval was used here primarily because producing overt responses during testing would not be possible or natural. Forming the mental image of the missing pieces resulted in a more accurate reproduction of the map later relative to restudying the map, indicating that covert retrieval improved visuospatial memory. Similarly, Kang (2010) examined the mnemonic benefits of covertly retrieving in a situation where an overt response would be difficult or time consuming. In Kang's experiments, subjects learned a set of Chinese characters and their English translations. Then, subjects either practiced covert retrieval of the Chinese characters by forming a mental image of the characters in response to the English form, or restudied the pairs across two blocks. On the final retention test, subjects were provided with the English words and were required to draw the Chinese characters. Across three experiments, Kang showed that covertly retrieving the Chinese characters resulted in superior final performance relative to restudying.

These experiments show that covert retrieval benefits retention (Orlando & Hayward, 1978; Carpenter & Pashler, 2007; Kang, 2010). Therefore, one might conclude that Tulving (1983) is correct—there are no differences between an overt response and simply bringing the information to mind. However, these experiments do not address whether covert retrieval benefits retention to the same degree as overt retrieval. It is possible that covert retrieval benefits retention relative to control conditions where retrieval is not practiced at all, but that overtly producing the information provides an extra benefit. To my knowledge, little literature has examined the relative benefits of covert retrieval and producing an overt response during retrieval practice on later retention. What literature does exist presents mixed findings, with some experiments finding that an overt response produces greater retention benefits relative to covertly bringing information to mind, and other literature finding there are no differences between the two.

Within the literature on adjunct questions, some researchers have studied the relative benefits of overtly responding to questions and covertly answering questions. Experiments on adjunct questions examine the effects of answering questions while reading textbook materials (see Anderson & Biddle, 1975). Answering adjunct questions while reading text material can facilitate comprehension and retention of that material, an effect related to the effects of retrieval practice (Roediger & Karpicke, 2006b). Some research has indicated that overtly responding to the adjunct questions results in greater comprehension and retention relative to covertly responding, whereas other research has shown no differences between the two types of responses (e.g., Michael & Maccoby, 1953; Kemp & Holland, 1966). This might suggest that overtly practicing retrieval may

result in greater retention relative to covertly practicing retrieval. For example, Michael and Maccoby (1953) presented a short film to classes of high school students. During the film, some students were instructed to answer adjunct questions either overtly or covertly during breaks in the film. On a final test administered after the film had ended, there were no performance differences between students who overtly responded to the questions and those who covertly responded to the questions, and this variable did not interact with any of the other variables in the experiment. Kemp and Holland (1966), on the other hand, compared teaching machine programs in programmed instruction—programs requiring students to answer questions or fill in blanks as they progress through the material—with either overt or covert responding. Kemp and Holland used a blackout procedure where they tested to see how much information could be removed from programmed instruction materials before errors in responses to the adjunct questions significantly increased (see also Holland & Kemp, 1965). This blackout procedure was used to indicate what proportion of the teaching machine materials contained content critical for comprehension. They found that when the responses to the adjunct questions were related to the critical content presented, there was an advantage of overt responding over covert responding.

Thus research within the adjunct questions literature presents mixed results regarding the effectiveness of covert responding on later retention. However, even with these mixed results, the general conclusion from this literature has been that an overt response produces superior comprehension and retention. In their review of the adjunct questions literature, Anderson and Biddle (1975) suggested that "if in one category are placed all of the experiments in which subjects were requested to make an explicit written

response to adjunct questions and in another category those studies in which either covert, mental answers were permitted or the procedure was ambiguous, it is apparent that both the direct and indirect effects of adjunct questions are more consistent when the subjects make an overt response" (p. 99). As I mentioned previously, when subjects are instructed to covertly retrieve information we do not know what exactly they covertly retrieve. The quality and accuracy of what is retrieved during covert retrieval may differ among these experiments, and this could be one reason for the discrepant results reported within this literature. If subjects were equally motivated to retrieve during both overt and covert responding, it is possible that the retention benefits of practicing covert retrieval would be consistent with the retention benefits of practicing overt retrieval.

Others have noted an overall advantage for producing an overt response to adjunct questions as well. When Robinson (1941) proposed the SQ3R procedure, he recommended that students recite overtly as opposed to mentally reviewing the information in order to improve retention. Robinson reasoned that "the more sensory channels used in learning, the more effective it is; in writing notes one provides visual and kinaesthetic (muscle) cues as well as verbal imagery in thinking about it" (p. 30), leading to the suggestion that "[writing an answer out] is more effective since it forces the reader actually to verbalize the answer whereas a mental review may often fool a reader into believing that a vague feeling of comprehension represents mastery" (p. 30). Robinson's suggestions fall closely in line with theories of embodied cognition, the idea that cognitive processes are rooted in the body's interactions with the world (see Wilson, 2002). These theories lead to the hypothesis that producing an overt response when retrieving information will benefit retention more than just covertly bringing information

to mind. Research on the production effect might also suggest that there is something special about producing an overt response. The production effect refers to the fact that producing a word out loud during study results in greater retention of that word relative to words that were only read silently during study, at least when within-subject, mixed list designs are used (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Although the overt response occurs during study in production effect experiments, the positive effect for spoken words over words read silently suggests that producing information overtly results in superior memory relative to covertly rehearsing information. It is quite possible that this effect might also occur during testing, resulting in superior retention on a delayed test from overt retrieval relative to covert retrieval.

Izawa (1976) also examined the relative benefits of overt and covert retrieval investigating verbalized overt test trials and covert test trials in a multitrial paired-associate learning experiment. Izawa's experiment did not ask questions regarding the relative retention benefits of overt and covert retrieval, but instead examined forgetting across trials and potentiation of learning on subsequent study trials as a result overt and covert retrieval (see Izawa, 1969; 1971). In her experiment, subjects learned nonsense syllables or nouns during study and cued recall test trials. Across multiple cycles of one study trial followed by five test trials, Izawa manipulated whether retrieval was overt (subjects verbally produced their responses) or covert (subjects silently recalled their responses). At the very end, subjects completed one final overt test. Results indicated that sometimes an overt response was superior to a covert response, and sometimes there were no differences. On the one hand, overt test trials reduced forgetting between trials during the learning phase relative to covert trials. On the other hand, both overt and covert test

trials resulted in equivalent test potentiation effects: subjects learned more from a study trial that followed a test trial (with either overt or covert responding) than they learned from a study trial that did not follow a test trial. Thus, both overt and covert retrieval trials potentiated learning during subsequent study trials, and therefore the final level of learning was equivalent after overt and covert test trials. These results join the mixed literature showing that sometimes covert and overt retrieval produce equivalent benefits, and other times they do not. However, Izawa (1976) did not use a testing effect design as in most of he literature.

Based on the literature reviewed here, it is unclear whether covert retrieval leads to a retention benefit comparable to that of overt retrieval. Some literature has found positive effects of covert retrieval practice (e.g., Kang, 2010), and Tulving (1983) suggests that practicing covert retrieval will produce retention results comparable to practicing overt retrieval. This hypothesis falls in line with Tulving's theory that retrieval and memory performance are two separate elements of episodic memory (Tulving, 1983, pp. 134-137), and that the act of retrieval affects later retention. Conversely, other literature suggests that there is something special about producing information overtly during retrieval, and that overt retrieval should lead to superior levels of retention relative to covert retrieval (Kemp & Holland, 1966; Anderson & Biddle, 1975; Robinson, 1941; MacLeod et al., 2010). Theories of embodied cognition might also suggest that some type of overt response may further enhance later retention (see Wilson, 2002). Therefore, the purpose of this series of experiments was to directly test whether an overt response during retrieval produces a superior benefit on retention relative to covert retrieval. In prior work, there was no guarantee that subjects were as motivated to retrieve during

covert retrieval as they were during overt retrieval, and this problem may have led to the mixed results reported in the literature. In this study, I employed conditions that helped to ensure subjects were motivated to retrieve even during covert retrieval. Therefore, my procedure will test the relative retention benefits of covert and overt retrieval under better control.

Experiment 1

Experiment 1 employed 3 conditions manipulated within subjects to address the issue of whether covert and overt retrieval produce equivalent testing effects. Subjects first studied a categorized list and then completed an initial test. During the initial test, subjects were cued to overtly retrieve items from some categories and covertly retrieve items from other categories. Importantly, the overt and covert conditions were as nearly equated as possible and the two types of trials were intermixed so that subjects would be motivated to retrieve during both the overt and covert retrieval trials. During the initial test, when subjects were cued with a category name they always began by thinking of the items from that category (i.e., by covertly retrieving the items). Then, subjects were either prompted to explicitly produce the items (the overt retrieval condition) or to continue thinking of the items (the covert retrieval condition). Therefore, when subjects were given a category cue they did not know whether or not they would need to produce the items until after they had spent time covertly retrieving the items. A third set of categories was not cued during the initial test (no test control condition). After a short delay, subjects completed final free recall and final cued recall tests to assess retention of the items.

Method

Subjects. Thirty-six subjects (22 female, ages 18-35, median age of 20) were recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or pay. Two subjects were removed and replaced because they did not follow testing instructions.

Design. The experiment consisted of three within-subject conditions: overt retrieval, covert retrieval, and no test. For the overt retrieval condition, subjects produced the studied items during the initial test by typing them into the computer. For the covert retrieval condition, subjects did not produce the studied items during the initial test. Instead, subjects in this condition only reported how many items were brought to mind. For the no test condition, subjects were not cued to recall items at all.

Materials. Materials consisted of categorized word lists. Items from 18 different categories were taken from the updated version of the Battig and Montague (1969) word norms (Van Overschelde, Rawson, & Dunlosky, 2004). Six items were drawn from each category for this experiment. The first few items from each category were not used to help reduce the influence of guessing on the tests (see Tulving & Pearlstone, 1966), so items 5 through 10 were used. The categories were divided into three sets of six categories, and each set was fully counterbalanced across each of the three conditions (overt retrieval, covert retrieval or no test). In addition, the categories used contained differing numbers of items (4, 5 and 6) so that when subjects were asked to covertly retrieve and report the number of items they recalled they would be less likely to rely on category size for responding. Each of the three sets of categories contained two categories with 4 items, two with 5 items, and two with 6 items so that the total number of items in

each condition was equated. For each subject the computer randomly determined whether each category in a given set would be 4, 5, or 6 items in length. When only 4 or 5 items were to be studied, the computer randomly determined which of the six items were to be dropped from the presentation.

Procedure. The experiment began with a study phase, and all study items were presented visually on the computer screen. Each list was organized by category, and categories assigned to each of the three sets were evenly distributed throughout the study phase. The category name always appeared first for 2 seconds, and then the words from that category were presented one at a time for 2 seconds each with a 30-second interstimulus interval separating each presentation. In addition, category names were presented in all uppercase letters to indicate clearly to the subjects that a new category of words was beginning. Subjects were instructed to study the words as they appeared so that they would be able to recall them later.

After the study period, all subjects completed a filler task (playing Pac Man) for 3 minutes. After the filler task, subjects completed the initial test. During the initial test, all subjects were warned against guessing, and were told that the experimenter may ask them to recall the items again later in the experiment to encourage subjects to do their best during the initial test. Overt and covert retrieval trials were intermixed during the initial test. Subjects always began by thinking of the studied items belonging to the cued category during both the overt and covert retrieval trials for 40 seconds. In the overt condition, subjects were then instructed to continue thinking of the studied items from the cued category, but to type all of the items they could recall into the computer for 20 seconds. Finally, subjects were asked to type in the total number of studied items that

they had recalled from the cued category using a single digit from the numeric keypad. During the covert retrieval trials, subjects began with the same 40-second thinking phase. Then for the remaining 20 seconds, they were instructed to continue thinking of the items, but not to type them. Then, subjects were asked to type in the total number of studied items that they recalled from the cued category. Therefore, the initial 40-second thinking phase was the same for categories in both the overt retrieval and covert retrieval conditions. When subjects were presented with a category name and instructed to think of the studied items from that category they did not know whether they would be required to produce the items or just the number of items until after they had already attempted covert retrieval. This procedure should have helped to motivate subjects to covertly retrieve when instructed to do so. Finally, a third set of categories was not presented to the subjects at all during the initial test (the no test condition).

After subjects completed the initial test, they played Tetris for 15 minutes. Then, all subjects completed a final free recall test and a final cued recall test. First, subjects were asked to recall all of the studied words they could remember for 10 minutes. They were told to type as many studied items from as many categories as possible, but were also warned against guessing. Second, subjects were given a cued recall packet containing each category name from the study phase with six blank lines printed below the category names. They were instructed to recall as many items from each category as possible and were again warned against guessing. They were also told they could recall the items in any order they wanted. After subjects completed the cued recall test, they were debriefed and thanked for their time.

Results

All results were reliable at the .05 level of confidence unless otherwise noted. A Greenhouse-Geisser correction was used for violations of the sphericity assumption in repeated measures analyses (Geisser & Greenhouse, 1958). The Bonferroni correction for multiple comparisons was used for all pairwise comparisons unless otherwise noted. Error bars for all figures represent 95% confidence intervals corrected for within-subject designs (Cousineau, 2005; Morey, 2008).

Initial Test Performance. The results from the initial test are shown in Figure 1. As would be expected, the number of items produced (either overtly or covertly) during the initial tests was nearly identical. Subjects reported recalling, on average, the same number of items using the numeric keypad during the initial overt trials ($M = 2.93$, mean correct recall was 2.50, with an average of 0.43 intrusions per category) as they did during initial covert trials ($M = 2.94$; $t < 1$).

Final Free Recall Performance. Given that subjects reported retrieving the same number of items during each type of trial on the initial test, the next question is whether each type of retrieval produced the same benefit on retention. The results from the final free recall test are shown in Figure 2. A one-way ANOVA ($F(2,70) = 27.40$, $\eta_p^2 = .44$) indicated there were differences in final recall between conditions. Pairwise comparisons revealed that equivalent proportions of items were recalled from the overtly tested categories ($M = .45$) and the covertly tested categories ($M = .47$), and recall of these items was significantly better than recall of items assigned to the no test condition ($M = .26$). Thus, it appears that covert retrieval improves retention of verbal materials just as much

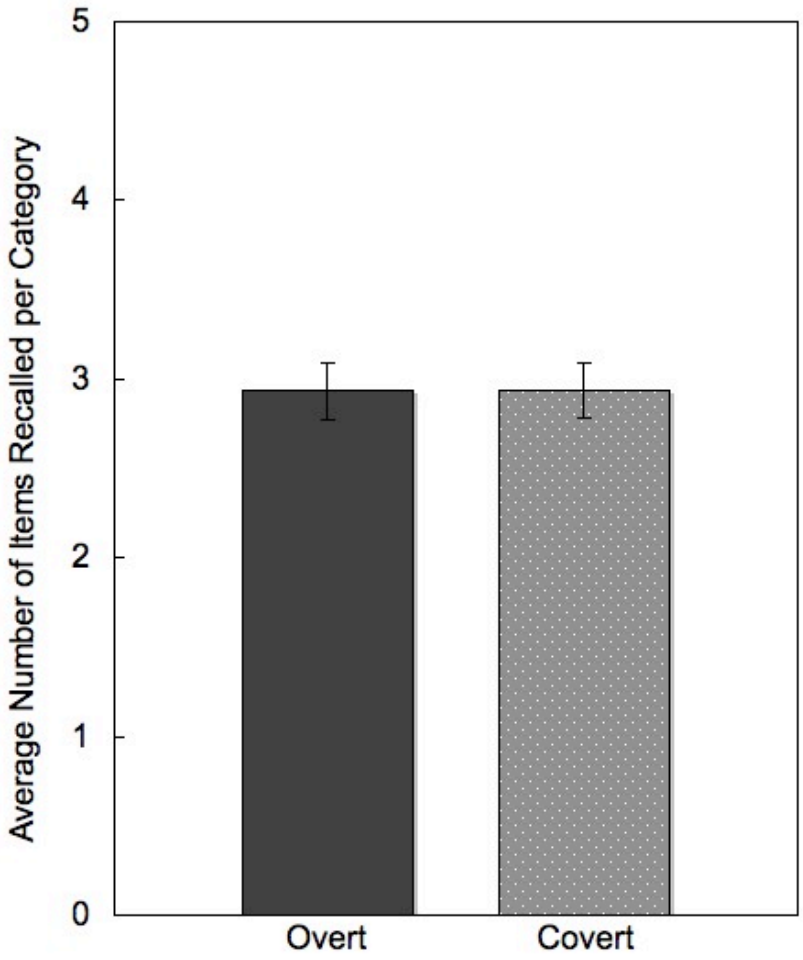

Figure 1. Average number of items produced per category on the initial tests for the overt and covert retrieval conditions in Experiment 1. Error bars represent within-subject 95% confidence intervals.

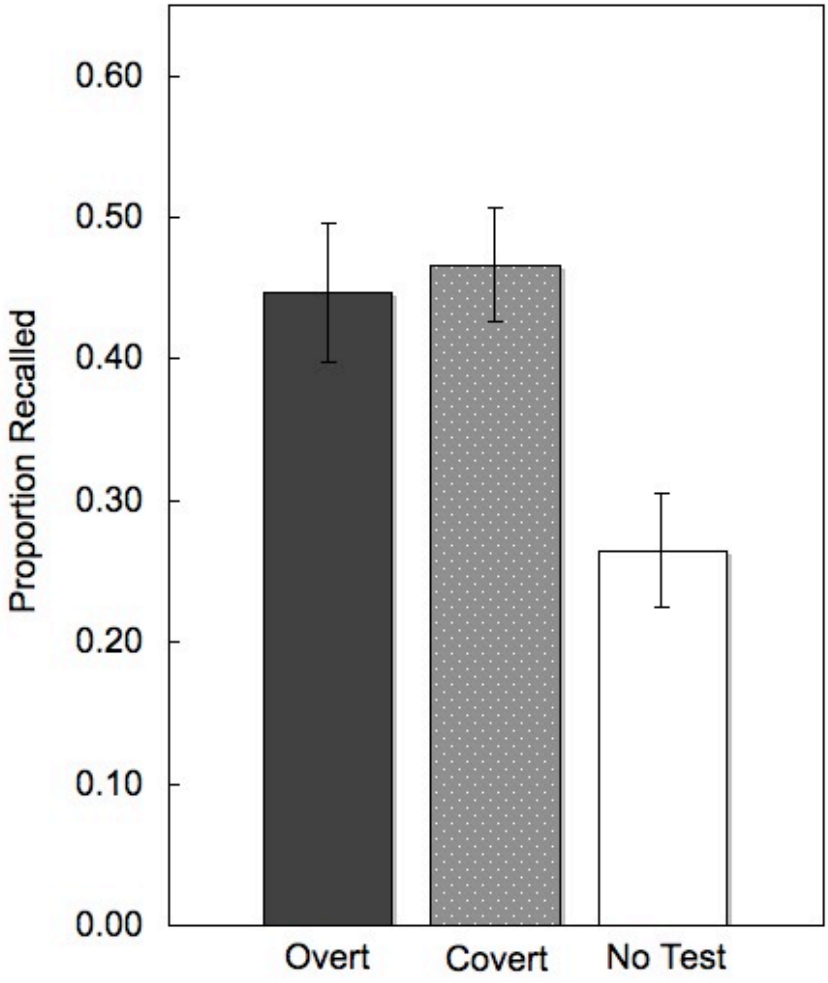

Figure 2. Performance on the final free recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 1. Error bars represent within-subject 95% confidence intervals.

as overt retrieval does, so an overt response is not necessarily needed to improve retention.

The number of intrusions produced per category on the final free recall test from overtly tested categories, covertly tested categories, and non-tested categories were also statistically analyzed. Of all intrusions produced, only 1.4% could not be categorized into one of the 18 studied categories and so these were left out of the analysis. There were no differences in the number of intrusions produced on the final free recall test across conditions ($F(2,70) = 1.07$) with means of 0.20, 0.25, and 0.18 per category for the overt, covert, and non-tested categories respectively. Thus, covertly retrieving the items from the list did not result in greater retention of incorrect information relative to overt retrieval.

The final free recall data were also analyzed using category recall and words-per-category recall (Tulving & Pearlstone, 1966), and these measures are shown in Table 1. Category recall measures the number of categories from which the subject recalls at least one word (Cohen, 1963). In contrast words-per-category recall measures how many words on average are recalled within a category once at least one word has been recalled from that category. The product of the category recall measure and words-per-category recall measure equals the total number of words recalled. Across conditions there were differences in category recall ($F(2, 70) = 44.86$, $\eta_p^2 = .56$). Subjects recalled more categories from the overt ($M = 4.42$) and covert retrieval conditions ($M = 4.69$) than from the no test condition ($M = 2.58$). There were no category recall differences between the overt and covert retrieval conditions. In addition, no differences were found among conditions on the word-per-category recall measure ($F(2, 70) = 1.95$). These analyses

Table 1

Measures of category recall and words-per-category recall on the final free recall test in Experiment 1.

	Category Recall	Words-Per-Category	Total Recall
Overt	4.42 (.22)	2.95 (.13)	13.42 (1.02)
Covert	4.69 (.19)	2.89 (.15)	14.00 (1.00)
No Test	2.58 (.25)	2.57 (.24)	7.94 (.92)

Note. Standard errors are reported in parentheses.

indicate that the superior performance of the retrieval conditions over the no test condition is driven by the ability to recall studied categories, and that overt and covert retrieval are similar in this respect.

Final Cued Recall Performance. Performance on the final cued recall test is shown in Figure 3. Performance on the final cued recall test was generally greater than performance on the final free recall test. This is to be expected because the category names should have been effective retrieval cues for the subjects particularly because the items were blocked by category. A one-way ANOVA revealed significant differences among the initial test conditions on the final cued recall test ($F(2,70) = 7.76$, $\eta_p^2 = .18$). Least significant difference post-hoc analyses indicated that there were no significant performance differences between the categories tested overtly ($M = .50$) and categories tested covertly ($M = .51$). Items from both the overtly and covertly recalled categories were remembered better than items from the non-tested categories ($M = .43$). There were again no differences in the number of intrusions produced per category from categories in each of the three conditions ($F < 1$) with means of 0.37, 0.36, and 0.38 for the overt, covert, and non-tested categories respectively. Thus the same pattern of results was obtained on the final cued recall test. It is of course possible that taking the final free recall test affected performance on the final cued recall test because the cued recall test always followed the free recall test (see Tulving & Pearlstone, 1966).

Discussion

In Experiment 1, I tested whether covert retrieval produces comparable retention benefits as overt retrieval by instructing subjects to overtly retrieve some items and covertly retrieve other items from a categorized list. Importantly, the overt and covert

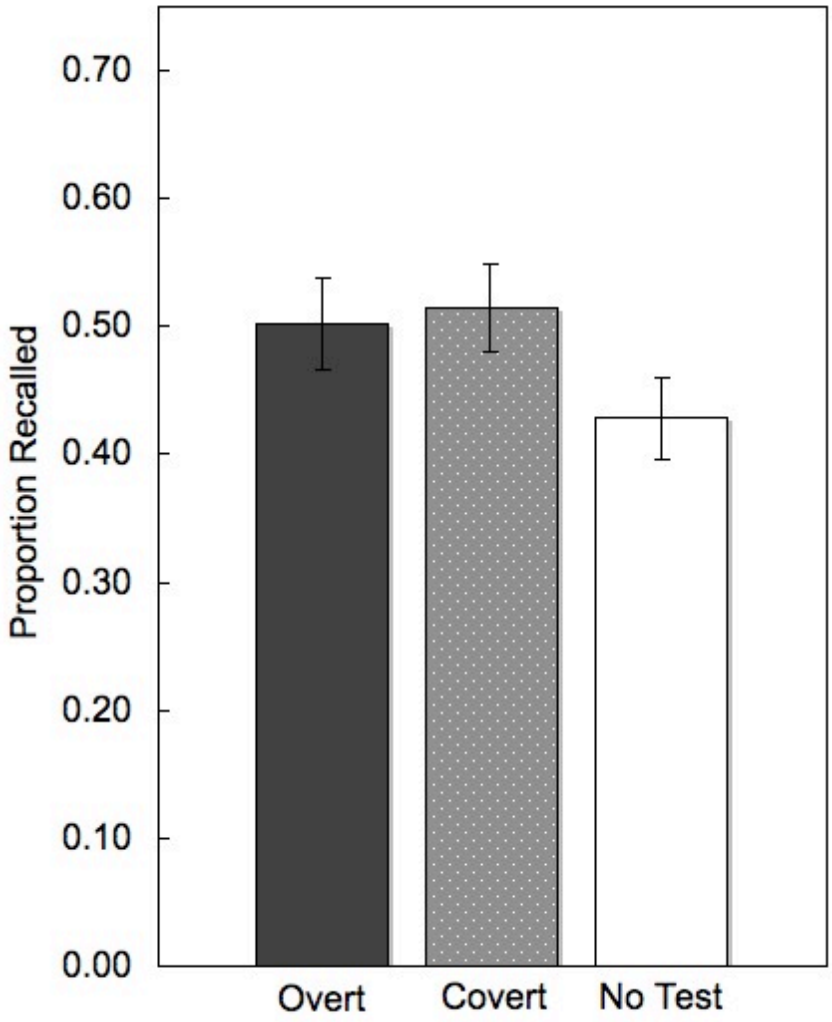

Figure 3. Performance on the cued recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 1. Error bars represent within-subject 95% confidence intervals.

retrieval conditions were intermixed and nearly equated to motivate subjects to retrieve during both the overt and covert retrieval trials. Results indicated that covert retrieval is just as effective at enhancing retention as overt retrieval. However, to make the overt and covert retrieval conditions so nearly equated, both conditions began with covert retrieval practice. It is possible that this procedure affected the way subjects retrieved during the overt retrieval condition making it less natural as compared to standard overt retrieval in most situations. When people normally retrieve information, they are thought to first bring the information to mind (i.e., they have a recollective experience), and then produce an overt response rather quickly thereafter (i.e., memory performance, see Tulving, 1983, pp. 134-137). However, in the overt retrieval condition of Experiment 1, I artificially forced subjects to covertly retrieve category members for a block of time before they produced overt responses. This procedure may have undermined possible benefits from overt retrieval because this aspect of the procedure was unnatural. Would allowing subjects to retrieve more naturally in the overt retrieval condition result in a larger testing effect for the overt relative to the covert retrieval condition? To answer this question, Experiment 2 was conducted to replicate the results from Experiment 1 using discrete overt and covert retrieval phases so that subjects could retrieve the items in each category more naturally, especially in the overt retrieval condition.

Experiment 2

Experiment 2 utilized the same 3 within-subjects conditions to address whether covert and overt retrieval produces equivalent testing effects. During the initial testing phase, subjects completed two distinct tests. During one initial test, subjects were cued with category names and were instructed to overtly retrieve the items. During the other

test subjects were instructed only to covertly retrieve the items. A third set of categories was not cued during the initial test (no test control condition).

Method

Subjects, Materials, and Design. Thirty-six subjects (20 female, ages 18-30, median age of 20) were recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or pay. None of the subjects had participated in Experiment 1. Two subjects were removed and replaced because they did not follow testing instructions. The materials and design for Experiment 2 were generally the same as in Experiment 1 with changes in the initial test to make the overt and covert retrieval conditions distinct from one another.

Procedure. The procedure for Experiment 2 was similar to that of Experiment 1. The experiment began with the same learning phase and the same filler task; however, the subjects participating in Experiment 2 completed two distinct initial tests, one for overt retrieval and one for covert retrieval (counterbalanced for order). During each initial test, subjects were presented with the category names assigned to the appropriate condition one at a time for 60 seconds. During the 60 seconds, subjects were instructed to recall as many studied items from the presented category as they could remember. Once again, subjects were warned against guessing and told that the experimenter may ask for recall of the items again later. During the overt initial test, subjects were instructed to type the studied items belonging to the category as they recalled them. During the covert initial test, subjects were instructed to think of the items belonging to the category. Instead of typing the items during recall, subjects were instructed to type an X every time they covertly retrieved a studied item, and were asked to only type one X for each

individual item recalled. Importantly, they never typed the specific items during the covert initial test. Again, a third set of categories was not presented to the subjects during the initial test (the no test condition).

After the initial test, subjects played Tetris for 15 minutes as in Experiment 1. Then, they completed the same final free recall and final cued recall tests as those in Experiment 1 did.

Results and Discussion

Initial Test Performance. The results from the initial tests are shown in Figure 4. As in Experiment 1, the number of items produced (either overtly or covertly) during the initial tests was nearly identical. Subjects produced, on average, the same number of items per category during the overt test (writing out the words, $M = 3.17$, mean correct recall was 2.67 with an average of 0.50 intrusions per category) and the covert test (using a key-press response, $M = 3.21$; $t < 1$). The same pattern of results was obtained regardless of the order in which the initial tests were taken.

Final Free Recall Performance. The results from the final free recall test are shown in Figure 5. Once again, a one-way ANOVA indicated that there were differences among the initial test conditions ($F(2,70) = 24.79$, $\eta_p^2 = .42$). Pairwise comparisons indicated that free recall of items from the overt retrieval ($M = .46$) and the covert retrieval conditions ($M = .44$) were not significantly different from one another. Free recall of items from the no test condition (M = .27) was significantly less than that from the two retrieval conditions. This pattern of results was the same regardless of the order in which the initial tests were taken.

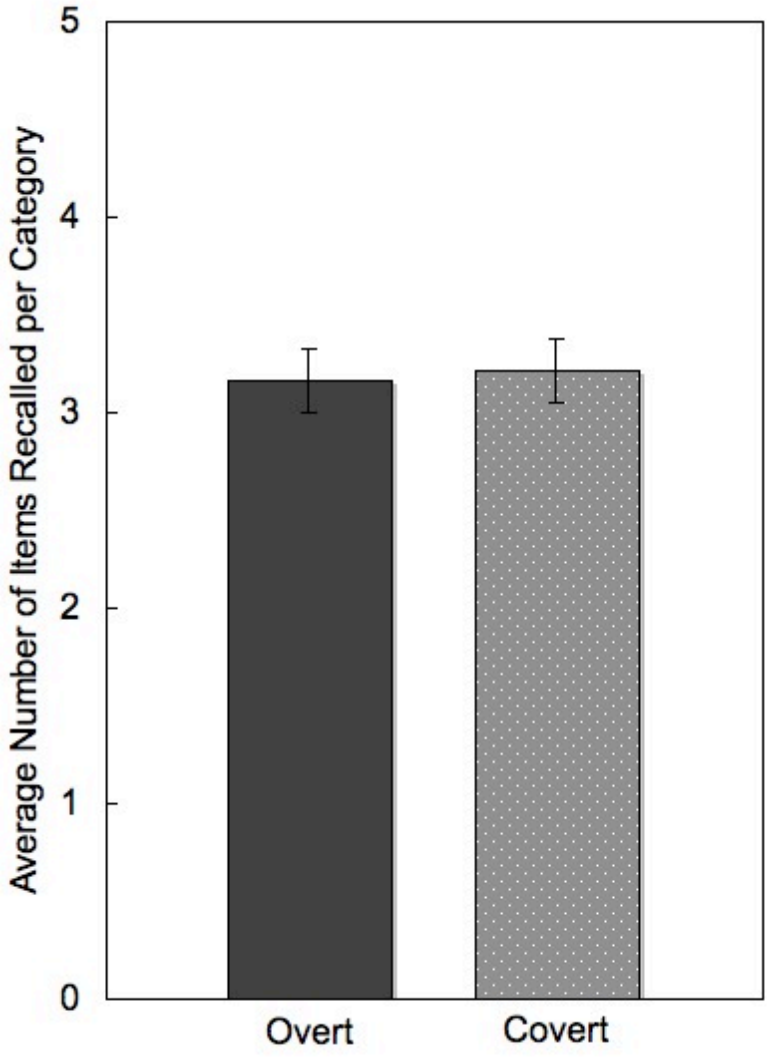

Figure 4. Average number of items produced per category on the initial tests for the overt and covert retrieval conditions for Experiment 2. Error bars represent within-subject 95% confidence intervals.

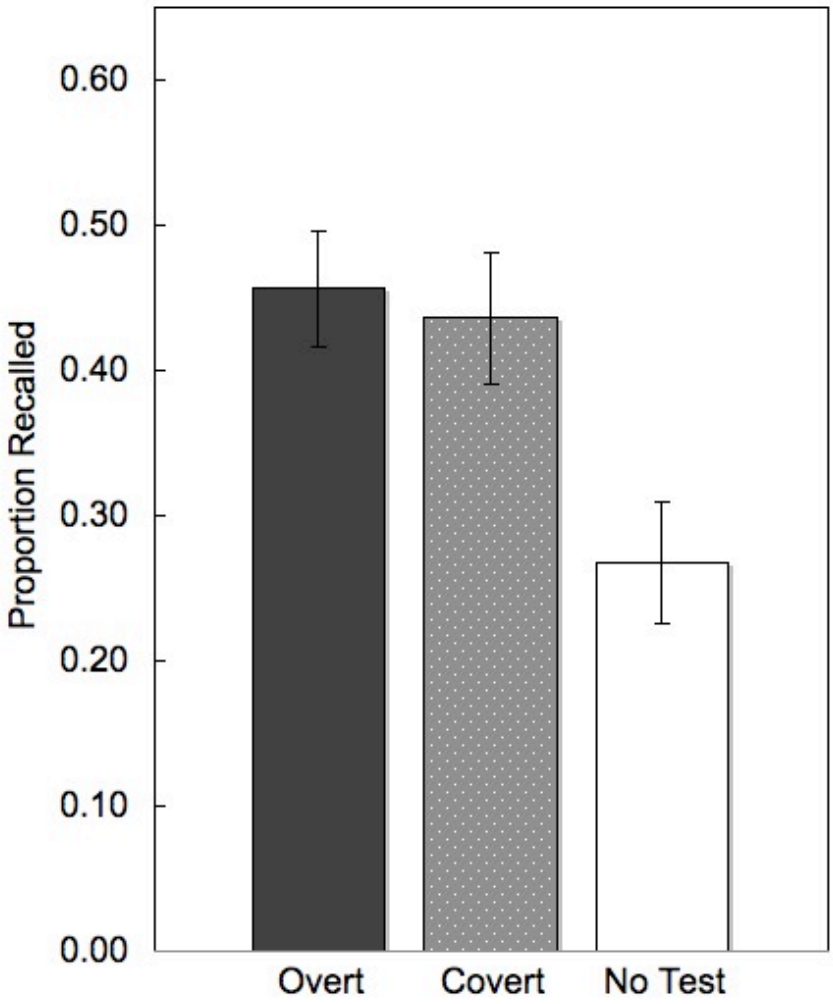

Figure 5. Performance on the final free recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 2. Error bars represent within-subject 95% confidence intervals.

In the same way as Experiment 1, intrusions produced on the final free recall test were analyzed across conditions. Of all intrusions produced, only 1.2% could not be categorized into one of the 18 studied categories and were left out of the analysis. In this experiment, there was a significant difference in intrusions produced per category between conditions ($F(2,70) = 5.17$, $\eta_p^2 = .13$). Post-hoc comparisons revealed that subjects produced significantly more intrusions per category from the overtly tested (M = 0.30) and covertly tested categories (M = 0.29) than from the non-tested categories (M = 0.16). Importantly, intrusions from the overt and covert conditions were not significantly different from each other.

Category recall and words-per-category recall (Cohen, 1963; Tulving & Pearlstone, 1966) from Experiment 2 are shown in Table 2. Once again there were significant differences between conditions for the category recall measure ($F(2,70)$ = 36.18, $\eta_p^2 = .51$). Subjects had higher category recall for the overt retrieval (M = 4.64) and covert retrieval (M = 4.36) conditions relative to category recall of the non-tested categories (M = 2.69). The overt and covert retrieval conditions were not different from one another. However, there were no differences between conditions for words-per-category recall ($F < 1$). These effects replicate those from Experiment 1, demonstrating that the recall was better for the retrieval conditions because subjects were able to recall more categories, and the overt and covert retrieval conditions yielded the same pattern of results. Therefore, Experiment 2 replicated the free recall results from Experiment 1: overt and covert retrieval resulted in comparable retention benefits.

Final Cued Recall Performance. The results from the final cued recall test are shown in Figure 6. Again, performance on the final cued recall test was generally greater

Table 2

Measures of category recall and words-per-category recall on the final free recall test in Experiment 2.

	Category Recall	Words-Per-Category	Total Recall
Overt	4.64 (.22)	2.83 (.16)	13.69 (1.07)
Covert	4.36 (.21)	2.90 (.13)	13.08 (.95)
No Test	2.69 (.23)	2.78 (.20)	8.03 (.87)

Note. Standard errors are reported in parentheses.

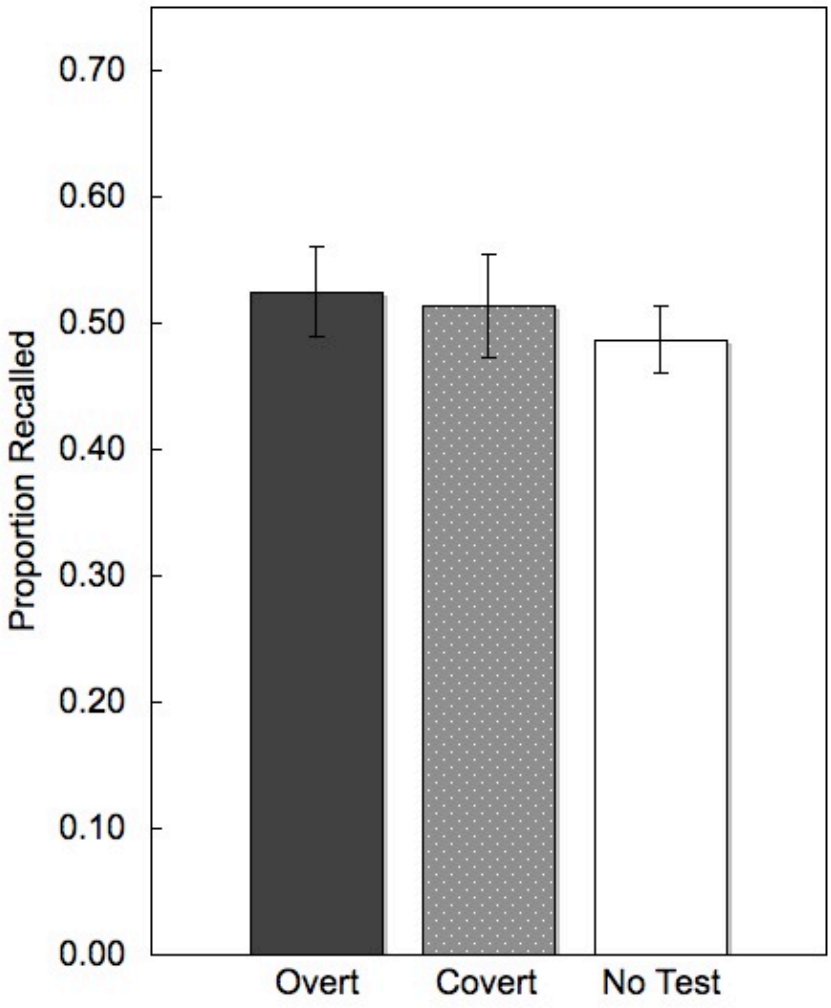

Figure 6. Performance on the cued free recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 2. Error bars represent within-subject 95% confidence intervals.

than performance on the final free recall test. However, there were no significant differences among the three conditions ($F(2,70) = 1.27$), and no significant differences in the number of intrusions produced ($F(2,70) = 1.44$). So, retention as measured by the cued recall test was the same for categories assigned to the overt and covert retrieval conditions. However, retention of the categories assigned to the retrieval conditions was not significantly better than retention for the not tested categories, although the means fall in the direction of greater recall for tested categories. Performance was slightly lower for the non-tested categories ($M = .49$) than the tested categories ($M = .52$), but the difference was small. Thus, no significant testing effect was found using the cued recall test as the final retention measure.

This result is different from that found in Experiment 1, and is not the usual finding (c.f., Zaromb & Roediger, 2010). Of course, the category name cues used on the cued recall test are very effective cues (Tulving & Pearlstone, 1966) and this may be one potential reason why there were no differences between conditions on this test. Analysis of the final free recall data (reported above) indicated that there were differences in the number of categories recalled between conditions, but no differences between conditions for the number of items per category recalled. Thus providing subjects with the category names during the cued recall test likely helped them to recall items from categories that were not previously recalled at all, possibly equating performance across the three conditions. This of course does not explain why there were different patterns of results between Experiment 2 and Experiment 1, where the same effective category name cues were used. It is possible that the cued recall results from one Experiment 2, which showed a trend towards greater recall from tested categories, was not powerful enough to

detect an effect. A t-test comparing cued recall between the two tested conditions combined and the no test condition yields a marginally significant result ($t(35) = 2.00$, $p = .05$). Nonetheless, across both experiments and both measures of final retention, it appears that covert retrieval enhances later retention just as much as overt retrieval does, at least when retention was measured relatively soon after the learning phase. Experiment 3 was carried out to examine the relative benefits of overt and covert retrieval on retention after a longer delay, and to compare the retrieval conditions to a restudy control condition.

Experiment 3

In Experiment 3 I asked whether covert retrieval would result in a retention benefit comparable to overt retrieval when retention was measured after a longer delay. One of the valued benefits of retrieval practice is that it improves long-term retention. It is possible that the effects of covert retrieval on retention are comparable to those of overt retrieval after a short delay, but not after a relatively longer delay. Therefore, before asserting that covert retrieval results in retention benefits comparable to overt retrieval, it is important to measure retention after a longer delay. Accordingly, in Experiment 3 one group of subjects completed the final free and cued recall tests during the initial learning session as in the first two experiments, whereas another group of subjects completed the final retention tests after a 2-day delay. In addition, a fourth within-subjects initial learning condition, a restudy control, was added to Experiment 3. Roediger and Karpicke (2006a) showed that relative to restudying, practicing retrieval prevents forgetting (see too Zaromb & Roediger, 2010). In Experiment 3, I examined forgetting across a 2-day

delay after overt retrieval practice, covert retrieval practice, restudy and no intervening test.

Method

Subjects. Forty-eight subjects (30 female, ages 18-43, median age of 19.5) were recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or pay. None of the subjects had participated in Experiments 1 or 2. Four subjects were removed and replaced because they did not follow test instructions.

Materials. Sixteen of the categories from the first two experiments were used. In addition, a differing number of items per category (5 or 6) were presented so that when subjects were asked to covertly retrieve items they would be less likely to rely on category size for responding. When 5 items were studied, the computer randomly determined which of the six items were presented to each subject. The categories were divided into four sets of four categories.

Design. The experiment utilized a 4 (initial test condition) X 2 (retention interval) design. The initial test variable consisted of four within-subject conditions: overt retrieval, covert retrieval, restudy, and no test. The overt retrieval, covert retrieval, and no test conditions were the same as in Experiment 2. During the restudy condition, subjects were presented with the items in each category assigned to the restudy set one at a time. As in Experiment 2, the conditions were blocked during the initial phase. The order of the category sets was held constant, and the initial test conditions were fully counterbalanced across the sets. Retention interval was manipulated between subjects; some subjects completed the final tests immediately after the learning phase (the immediate condition),

and the other group returned 2 days later to complete the final tests (the delayed condition).

Procedure. Subjects began with a study phase that was identical to that of Experiment 2. Subjects then completed the filler task (they played Pac Man) for 3 minutes, and then completed each of the two initial tests and the restudy phases. The overt and covert retrieval tests were the same as in Experiment 2, except subjects were given 30 seconds to retrieve for each category cue. This was done so that the testing conditions and the restudy condition could be equated for time, and because subjects in the first two experiments reported that they had more time than was necessary to retrieve for each category cue. During the restudy phase, subjects restudied the four categories assigned to the restudy condition. As in the first study phase, the category name was presented first for 2 seconds in all uppercase letters followed by the items in each category. However, the interstimulus interval was lengthened so that the restudy phase took the same amount of time as the overt and covert initial tests.

After subjects finished the overt test, covert test, and the restudy phase, they completed a distraction phase for 15 minutes (they played Tetris). Then, subjects in the immediate retention interval condition completed the same final free recall and cued recall tests as in Experiment 2. Subjects in the delayed retention interval condition were dismissed and asked to return to the lab 2 days later. When they returned, they completed the same final free recall and cued recall tests. Once the final tests were complete, all subjects were debriefed and thanked for their time.

Results and Discussion

Initial Test Performance. The results from the initial tests are shown in Figure 7. Again, the number of items produced either overtly or covertly during the initial tests was nearly identical. A 2 (initial test condition) X 2 (retention interval) ANOVA indicated that subjects produced, on average, the same number of items per category during the overt test (writing out the words, $M = 3.24$; mean correct recall was 2.45 with an average of 0.79 intrusions per category) and the covert test (using a key-press response, $M = 3.06$; $F(1, 46) = 1.39$, *ns*). There were no performance differences between subjects in the immediate retention interval group ($M = 3.03$) and the delayed retention interval group ($M = 3.28$; $F(1, 46) = 1.02$, *ns*) and no interaction ($F(1, 46) = 1.39$, *ns*).

Final Free Recall Performance. The results from the final free recall test are shown in Figure 8. A 4 (initial test condition) X 2 (retention interval) ANOVA revealed that overall there were differences among the initial test conditions ($F(3, 138) = 29.46$, $\eta_p^2 = .39$), and forgetting occurred overall—subjects in the immediate group ($M = .35$) performed significantly better than those in the delayed group ($M = .22$; $F(1, 46) = 8.74$, $\eta_p^2 = .16$). However, these effects were qualified by a significant interaction ($F(3, 138) = 4.71$, $\eta_p^2 = .09$). The interaction revealed that restudying resulted in superior short-term retention relative to retrieving the items (either overtly or covertly) or doing nothing (the no test condition), but this advantage did not hold after a longer delay (see also Roediger & Karpicke, 2006a).

Post hoc analyses confirmed these observations. Subjects in the immediate retention interval group recalled significantly more items from the restudied categories ($M = .58$) than from categories overtly recalled ($M = .34$), covertly recalled ($M = .32$) and

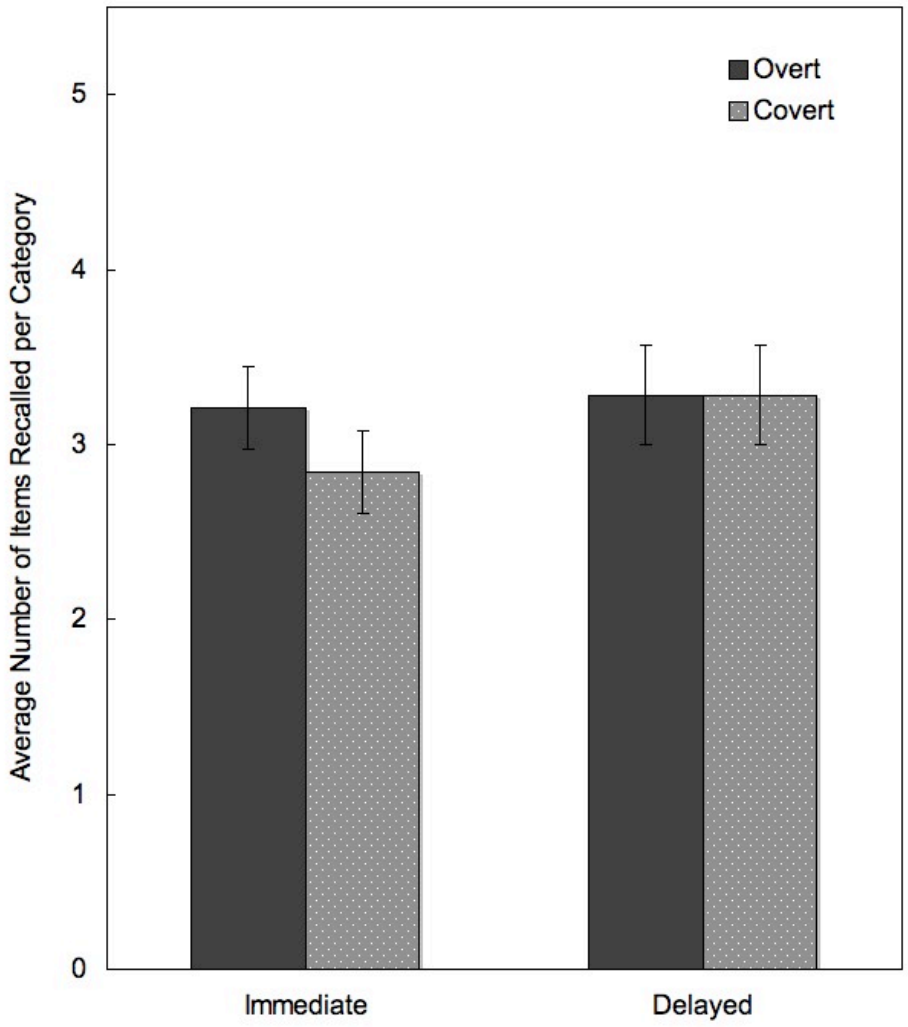

Figure 7. Average number of items produced per category on the initial tests for the overt and covert retrieval conditions for both the immediate and delayed retention interval conditions in Experiment 3. Error bars represent within-subject 95% confidence intervals.

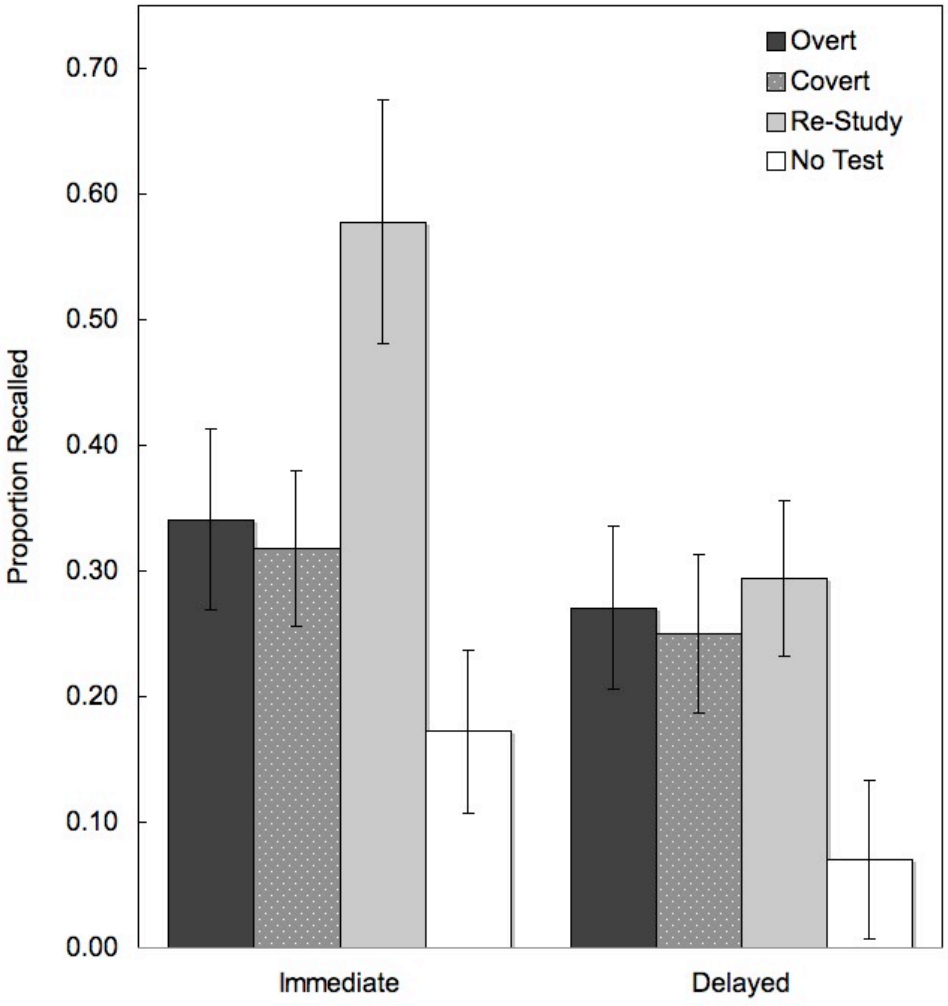

Figure 8. Performance on the final free recall test for the overt retrieval, covert retrieval, restudy, and no test conditions and for both the immediate and delayed retention interval conditions in Experiment 3. Error bars represent within-subject 95% confidence intervals.

those not tested (*M* = .17). In addition, these subjects recalled significantly fewer items from the non-tested categories relative to items from categories assigned to the other three conditions. Importantly, recall from the overtly tested categories and the covertly tested categories did not differ. A slightly different pattern of results was found for subjects in the delayed retention interval group. For these subjects, recall of items from the non-tested categories (*M* = .07) was significantly worse than recall from the overtly tested categories (*M* = .27), covertly tested categories (*M* = .25), and restudied categories (*M* = .29). No other comparisons reached significance. Thus it appears that restudying category members resulted in a great deal of forgetting (.58 vs .29 from the immediate to the delayed final free recall test), and practicing retrieval both overtly and covertly protected against this forgetting (.34 vs .27 for the overtly tested categories, and .32 vs .25 for the covertly tested categories).

Proportional measures of forgetting—(initial recall – final recall)/initial recall (Loftus, 1985)—also indicated that practicing retrieval either overtly or covertly protected against forgetting. The proportional measures of forgetting are shown in Figure 9. Practicing overt and covert retrieval only resulted in 21% forgetting using this measure. However, restudying the category members and doing nothing with the category members (the no test condition) resulted in much more forgetting than practicing retrieval did (49% and 59% forgetting, respectively). This finding demonstrates that practicing retrieval protected against forgetting, and it made no difference whether retrieval was performed overtly or covertly.

As in the first two experiments, intrusions produced per category on the final free recall test were analyzed across conditions. Of all intrusions produced 6% and 19% were

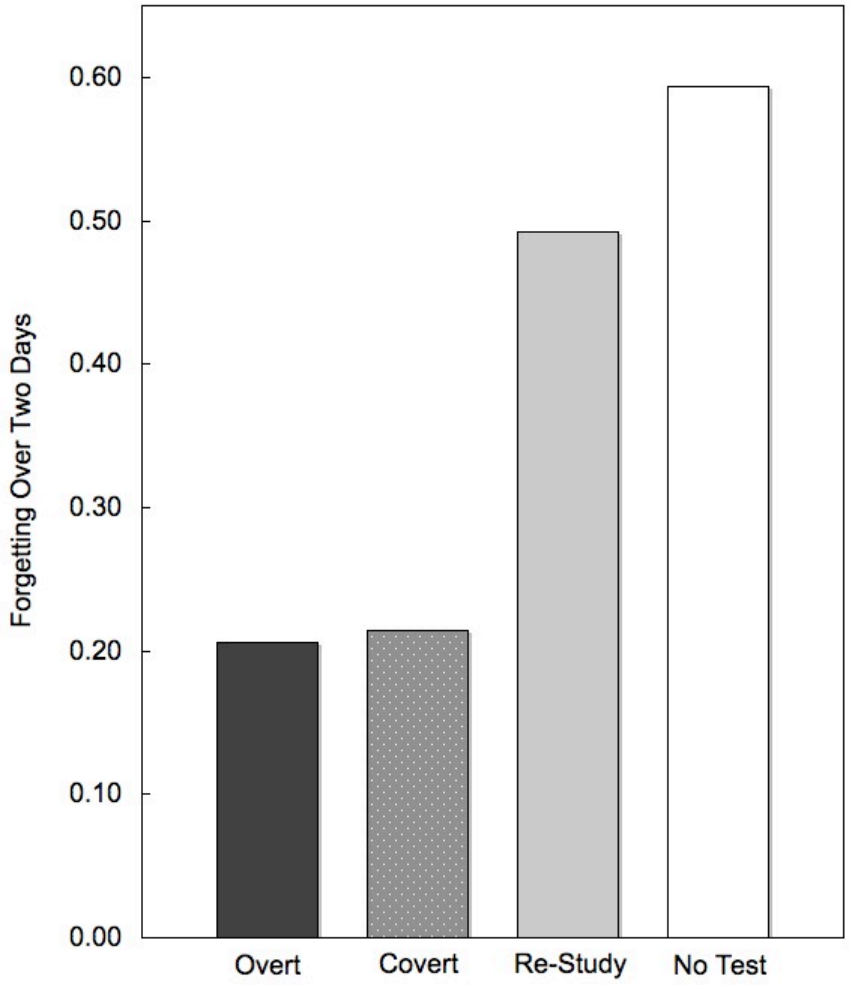

Figure 9. Forgetting over 2 days for the overt retrieval, covert retrieval, restudy, and no test conditions.

of this type for subjects in the immediate and delayed retention interval groups respectively. For the delayed group, this percentage of intrusions is higher than from the immediate group and from previous experiments. The larger number of intrusions that could not be categorized into one of the studied categories produced by the delayed group is understandable, however, due to the longer delay between the study phase and the final test. As in the previous experiments, these intrusions that could not be categorized into one of the 16 studied categories were not included in the analysis. A 4 (initial test condition) X 2 (retention interval) ANOVA revealed that overall there were differences among the initial test conditions ($F(3, 138) = 6.27$, $\eta_p^2 = .12$). Post hoc analyses indicated that subjects produced significantly more intrusions per category from the overtly ($M = 0.36$) and covertly ($M = 0.24$) recalled categories than from those categories not tested ($M = 0.13$). Intrusions per category produced from the restudied categories ($M = 0.23$) were not statistically different from the other categories, and no other comparisons reached significance. The main effect of retention interval ($F(1, 46) = 1.34$, $\eta_p^2 = .03$) and the interaction ($F(3, 138) = 2.22$, $p = .11$, $\eta_p^2 = .05$) did not reach significance. Most importantly, practicing covert retrieval did not result in differing levels of incorrect information on the final free recall test relative to practicing overt retrieval.

Category recall and words-per-category recall (Cohen, 1963; Tulving & Pearlstone, 1966) from Experiment 3 are shown in Table 3. A 4 (initial test condition) X 2 (retention interval) ANOVA on the category recall results revealed that overall there were differences among the initial test conditions ($F(3, 138) = 30.29$, $\eta_p^2 = .40$). Post hoc comparisons indicated that category recall was significantly lower for categories that were not tested ($M = 0.94$) relative to categories that were overtly tested ($M = 2.19$),

Table 3

Measures of category recall and words-per-category recall on the final free recall test in Experiment 3.

	Category Recall	Words-Per-Category	Total Recall
Immediate condition			
Overt	2.33 (.29)	2.73 (.33)	7.50 (1.08)
Covert	2.21 (.26)	2.53 (.32)	7.00 (1.12)
Restudy	3.25 (.16)	3.82 (.21)	12.71 (1.01)
No Test	1.25 (.21)	2.06 (.37)	3.79 (.75)
Delayed condition			
Overt	2.04 (.24)	2.26 (.29)	5.96 (1.01)
Covert	2.00 (.27)	2.27 (.26)	5.50 (.95)
Restudy	2.17 (.26)	2.44 (.27)	6.46 (.99)
No Test	0.63 (.16)	1.13 (.30)	1.54 (.46)

Note. Standard errors are reported in parentheses.

covertly tested (M = 2.10), and restudied (M = 2.71). In addition, category recall was higher for categories that were restudied (M = 2.71) relative to those that were covertly retrieved (M = 2.10). No other comparisons reached significance. There was also a main effect of retention interval ($F(1, 46) = 5.26$, $\eta_p^2 = .10$) indicating that forgetting occurred from the immediate to the delayed final tests. Subjects in the immediate group (M = 2.26) recalled significantly more categories than those in the delayed group (M = 1.71). The interaction only reached a marginal level of significance ($F(3, 138) = 2.14$, $p = .10$, $\eta_p^2 = .04$). Most importantly, as in the previous two experiments, category recall between the overtly and covertly tested categories did not differ.

Contrary to the results from the first two experiments, there were also differences found using the words-per-category recall measure. In Experiment 3, words-per-category recall showed the same results as category recall. A 4 (initial test condition) X 2 (retention interval) ANOVA on the words-per-category recall results revealed a significant main effect of initial condition ($F(3, 138) = 14.94$, $\eta_p^2 = .25$). Post hoc comparisons indicated that overall words-per-category recall was significantly lower for categories that were not tested (M = 1.59) relative to categories that were overtly tested (M = 2.50), covertly tested (M = 2.40), and restudied (M = 3.13). In addition, words-per-category recall was higher for categories that were restudied (M = 3.13) relative to those that were covertly retrieved (M = 2.40). No other comparisons reached significance. There was also a main effect of retention interval ($F(1, 46) = 6.20$, $\eta_p^2 = .12$) indicating that forgetting occurred from the immediate to the delayed final tests. Subjects in the immediate group (M = 2.79) recalled significantly more words-per-category than those in the delayed group (M = 2.02). The interaction only reached a marginal level of

significance ($F(3, 138) = 2.31$, $p = .08$, $\eta_p^2 = .05$). Most importantly, as in the previous two experiments, words-per-category recall between the overtly and covertly tested categories did not differ.

Final Cued Recall Performance. The results from the final cued recall test are shown in Figure 10. Performance on the final cued recall test was generally greater than performance on the final free recall test. These results also replicated the results from the final free recall test in Experiment 3. A 4 (initial test condition) X 2 (retention interval) ANOVA revealed that overall there were differences among the initial test conditions ($F(3, 138) = 31.88$, $\eta_p^2 = .41$), and subjects in the immediate group ($M = .50$) performed significantly better overall than those in the delayed group ($M = .32$; $F(1, 46) = 9.01$, $\eta_p^2 = .16$). As with the free recall data, these results were qualified by a significant interaction ($F(3, 138) = 6.03$, $\eta_p^2 = .12$). Subjects in the immediate retention interval group recalled significantly more category members from the restudied categories ($M = .65$) than from the overtly tested categories ($M = .45$), covertly tested categories ($M = .43$), and non-tested categories ($M = .34$). In addition, these subjects recalled significantly fewer category members from the non-tested categories than the overtly tested categories. No other comparisons reached significance. On the other hand, subjects in the delayed retention interval group recalled significantly fewer category members from the non-tested categories ($M = .17$) than from the overt categories ($M = .36$), covert categories ($M = .36$), and restudied categories ($M = .38$). No other comparisons reached significance. Thus, as with the free recall data, restudying the category members resulted in a short-term advantage relative to practicing retrieval (either overt or covert) and doing nothing (the no test condition), but this advantage did not hold after a longer delay. Most

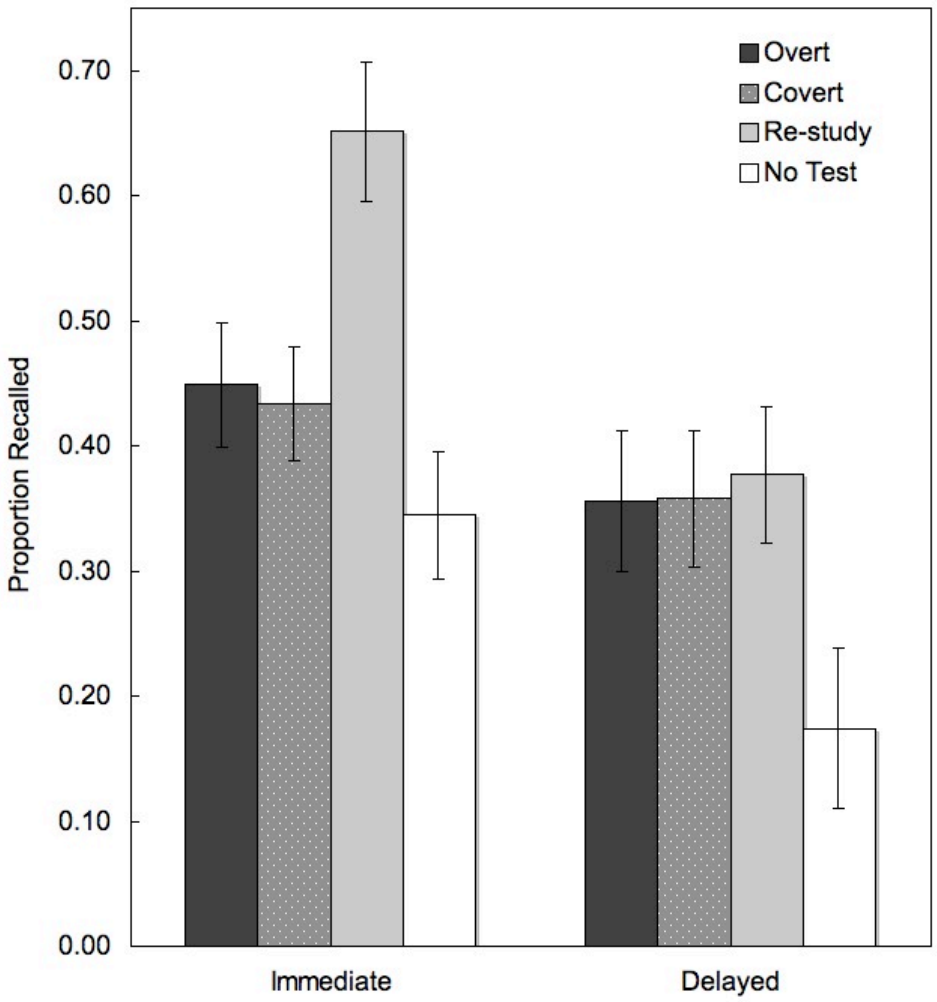

Figure 10. Performance on the cued free recall test for the overt retrieval, covert retrieval, restudy, and no test conditions and for both the immediate and delayed retention interval conditions in Experiment 3. Error bars represent within-subject 95% confidence intervals.

importantly, on both the immediate and delayed cued recall tests performance did not differ after subjects overtly retrieved category members and after they covertly retrieved category members.

A 4 (initial test condition) X 2 (retention interval) ANOVA was performed on the intrusions produced per category during the final cued recall test as well. The main effect of initial test condition ($F(3, 138) = 1.18$, $\eta_p^2 = .03$) and retention interval ($F(1, 46) = 1.54$, $\eta_p^2 = .03$) did not reach significance. However, there was a significant interaction ($F(3, 138) = 3.24$, $\eta_p^2 = .07$). The interaction revealed that subjects in the immediate retention interval group produced significantly fewer intrusions per category from restudied categories ($M = 0.23$) relative to overtly tested categories ($M = 0.49$), covertly tested categories ($M = 0.28$), and those categories not tested ($M = 0.48$). However, for subjects in the delayed group there were no significant differences in the production of intrusions per cagegory among the initial testing conditions (mean intrusions produced from the overt, covert, restudied and non-tested categories were 0.53, 0.56, 0.56, and 0.47 respectively). Once again, subjects did not produce differing amounts of intrusions per category after overt and covert retrieval on either the immediate cued recall test or the delayed cued recall test.

General Discussion

In the first two experiments, taking an initial test resulted in superior recall of the categorized word lists 15 minutes later relative to a no test control. However, later retention was the same when subjects practiced overt retrieval and when they practiced covert retrieval during the initial tests. In Experiment 3, I replicated these results on a final free recall test completed 15 minutes after learning and 2 days after learning—more

items were recalled from the overt and covert categories than the no test categories, and there were no differences between overt and covert retrieval conditions. In addition, restudying the items during initial learning resulted in the best performance on the immediate final free recall test, but resulted in the most forgetting after the 2-day delay. Practicing retrieval either overtly or covertly prevented a large amount of this forgetting across the delay. However, practicing retrieval did not result in superior absolute performance relative to restudying on the delayed final test, and this result is unlike the results from some other papers (c.f., Roediger & Karpicke, 2006a). This could have occurred because feedback was not provided after the initial tests in my experiment, and subjects were not given the opportunity to practice repeated retrieval of the items. Perhaps adding these features to the design would result in superior absolute performance relative to a restudy control on a delayed free recall test.

The results from the final free recall test were generally replicated with the final cued recall test in Experiment 1 and Experiment 3. In Experiment 2, there were no differences among the overt, covert, and no test conditions on the final cued recall test. This could have occurred because the category names provided as cues on the test should have been particularly helpful (Tulving & Pearlstone, 1966). On the free recall test retention differences after practicing retrieval (overtly or covertly) and doing nothing (the no test condition) were driven by category recall and not words-per-category recall, and providing the category names removed the burden of category recall during the cued recall test. Since there were no differences in words-per-category recall on the final free recall test after practicing retrieval or doing nothing initially, one might not expect to see differences on the final cued recall test. However, this does not explain why the results

differed between Experiments 1 and 3 and Experiment 2. It is also possible that the results from the cued recall test in Experiment 2 were simply obtained by chance. Still the critical result was obtained across all three experiments: retention was the same between categories that were overtly retrieved initially and categories that were covertly retrieved initially.

The fact that performance on the initial covert test cannot be scored may seem like cause for concern. In fact, this is probably the primary reason that covert retrieval is not frequently employed in experiments examining the effects retrieval practice via testing on later retention. In the experiments reported here there is no way to know exactly what the subjects retrieved during the initial covert test. Even though subjects reported retrieving the same number of items per category, a critic could perhaps argue that retrieval during the covert and overt tests might have differed, and subjects simply reported recalling the same number of items. However, this possibility is unlikely because the number of items per category varied. It would have been very difficult for subjects to rely on the size of the categories during covert retrieval for responding. Another critic might argue that intrusion rates could have differed between the two types of initial tests causing performance on the initial tests to only appear the same. This is also unlikely for a couple of reasons. On the final free and cued recall tests, there were no differences in intrusion rates between overtly tested categories and covertly tested categories, suggesting that intrusion rates also did not differ during the initial tests. In addition, the experiment conducted by Izawa (1976) described in the introduction also suggests intrusion rates do not differ between overt and covert retrieval attempts. When Izawa manipulated the number of overt (verbal) and covert (silent) test trials in her

multitrial learning experiment, intrusion rates during learning did not differ among subjects that completed all overt test trials and subjects that completed primarily covert test trials and only a few overt trials. Based on this result, Izawa suggested that subjects' wrong guesses did not differ for the overt and covert tests. Given these considerations it seems unlikely that subjects retrieved different amounts of correct information during the overt and covert initial tests.

Most importantly, the results from the three experiments reported here provide support for Tulving's (1983) hypothesis regarding overt and covert retrieval. Covertly retrieving information by bringing it to mind produces the same retention benefit as overt retrieval. Thus, even though the effects of retrieval practice have predominantly been studied using tests requiring subjects to make overt responses, these overt responses are not necessary for retrieval to benefit retention. However, even given these results, in some cases it still may be desirable to require overt responding during testing in educational settings. Testing can be beneficial in many other ways that are relevant for education beyond just directly improving retention (see Roediger et al., in press). For some of these benefits it is possible that a test requiring overt retrieval might be better than a test only requiring covert retrieval. For example, testing can be used to provide feedback to instructors. Clearly a test employing covert retrieval practice cannot provide such feedback.

Testing can also help students to improve their metacognitive monitoring—how accurate they are at judging how well they know the material—relative to restudying. When students repeatedly restudy their materials they are often over confident, but testing does not cause such over confidence (e.g., Roediger & Karpicke, 2006a; Karpicke

& Roediger, 2008). Testing can also be used to identify what students know and do not know, and can guide further efficient study as well. In fact, when students use self-testing as a study strategy it is often in order to exploit this benefit (Karpicke, Butler, and Roediger, 2009; Karpicke, 2009). Of course experiments demonstrating the metacognitive benefits of testing have required overt responses during testing. It is possible that covert retrieval practice may not help students to identify gaps in knowledge and improve metacognitive monitoring as well as overt retrieval practice. Although I do not know of any literature that has tested this idea empirically, Robinson (1941) recommends that students recite their lessons overtly rather than covertly for the purposes of diagnosing the state of one's knowledge in his book about effective study strategies: "Self-recitation may consist of mentally reviewing the answer or writing it out. The latter is more effective since it forces the reader actually to verbalize the answer whereas a mental review may often fool a reader into believing that a vague feeling of comprehension represents mastery" (p. 30). Further research will be needed to determine whether overt and covert retrieval practice affect students' metacognitions in the same way.

Overall, the results of the three experiments reported here indicate that covert retrieval practice works to enhance retention just as well as overt retrieval. Employing covert retrieval instead of overt retrieval in educational settings may save classroom time—teachers can simply pose questions to their students instead of writing test questions and scoring these questions later. Thus, as long as educators can be sure students are retrieving the answers, educators can implement activities in the classroom

that require retrieval practice without worrying about whether students produce an overt response if their goal is to improve their students' retention of the material.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159-177.

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9, pp. 90-132). New York: Academic Press.

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

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,* 941-956.

Butler, A. C., & Roediger, H. L. (2008). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 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,* 268-276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14,* 474-478.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633-642.

Cohen, B. H. (1963). An investigation of recoding in free recall. *Journal of Experimental Psychology, 65,* 368-376.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology, 1,* 42-45.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F* distribution in multivariate analysis. *Annals of Mathematical Statistics, 29,* 885-891.

Holland, J. G., & Kemp, F. D. (1965). A measure of programming in teaching-machine material. *Journal of Educational Psychology, 56,* 264-269.

Izawa, C. (1969). Comparison of reinforcement and test trials in paired-associate learning. *Journal of Experimental Psychology, 81,* 600-603.

Izawa, C. (1971). The test-trial potentiating model. *Journal of Mathematical Psychology, 8,*, 200-224.

Izawa, C. (1976). Vocalized and silent tests in paired-associate learning. *American Journal of Psychology, 89,* 681-693.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441-1451.

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38,* 1009-1017.

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

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469-486.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471-479.

Karpicke, J. D. & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966-968.

Kemp, F. D., & Holland, J. G. (1966). Blackout ratio and overt responses in programed instruction: Resolution of disparate results. *Journal of Experimental Psychology, 57,* 109-114.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 397-406.

Macleod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 671-685.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (in press). Test-enhanced learning in a middle school science classroom: The

effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399-414.

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,* 494-513.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20,* 516-522.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200-206.

Michael, D. N., & Maccoby, N. (1953). Factors influencing verbal learning from films under varying conditions of audience participation. *Journal of Experimental Psychology, 46,* 411-418.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4,* 61-64.

Orlando, V. P., & Hayward, K. G. (1974). A comparison of the effectiveness of three study techniques for college students. In P. D. Peterson & J. Hansen (Eds.), *Reading: Disciplined inquiry in process and practice* (pp. 242-245). Clemson, SC: National Reading Conference.

Robinson, F. P. (1941). *Effective Study*. New York: Harper and Brothers.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B.

Wright (Eds.), *New frontiers in applied memory* (pp. 13-49). Brighton, U.K.: Psychology Press.

Roediger, H. L., & Karpicke, J. D. (2006a). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

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

Roediger, H. L., Putnam, A. L., & Smith, M. A. (in press). Ten benefits of testing and their applications to educational practice. Chapter to appear in J. P. Mestre & B. H. Ross (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory.* Oxford: Elsevier.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Tulving, E. & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381-391.

Van Overshelde, J., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An expanded and updated version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289-334.

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9,* 625-636.

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