1-1-2011

The Effects of Response Modality on Retrieval

Adam Putnam

WASHINGTON UNIVERSITY

Department of Psychology

THE EFFECTS OF RESPONSE MODALITY ON RETRIEVAL

by

Adam Lewis Putnam

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

August 2011

Saint Louis, Missouri

Acknowledgements

I would like to thank the members of my Thesis Examination Committee, Dr. Henry L. Roediger, III, Dr. David A. Balota, and Dr. Mitchell S. Sommers for their support and advice throughout this research. I also thank Jinkun Zhang, Benjamin Hoener, Paige Madara, Kate Margolis, and Kelly Young for their help in collecting data. This research was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation.

Table of Contents

List of Tables

List of Figures

Abstract

The testing effect refers to the finding that retrieval practice can lead to enhanced recall on future tests. Despite being a widely researched phenomenon, the underlying mechanisms of the testing effect remain unknown, and basic issues are unresolved. The purpose of these experiments was to investigate how different response modalities influence retrieval both on initial and delayed tests. More specifically, we were interested in whether subjects can recall more via writing or speaking, whether writing (or speaking) on a first test can lead to better performance on a second test (and whether the type of second test would matter), and whether any form of overt retrieval on a first test leads to better performance on a final test compared to just thinking about a response. All of these questions were aimed at determining whether the beneficial effects of testing arise from the act of retrieval or are somehow tied to the production of the answer. Three experiments show that there are only small, if any, differences between typing and speaking performance, and that an initial covert retrieval will often yield the same benefit to future test performance as retrieval with an overt response production. The practical implications for education suggest that in rehearsing information, just thinking about an answer is just as beneficial to future retrieval as reporting answers aloud or writing them down.

The Effects of Response Modality on Retrieval

Many teachers and students develop theories about what makes something easy to remember. It would not be unusual to overhear a student reading her notes aloud as she prepares for a test, or for a teacher to suggest that writing a fact down may help solidify it in memory. People seem to have an intuitive feeling that writing, speaking, and reading information each influence memory in different ways.

In his book *The Elements of Episodic Memory,* Endel Tulving suggested that covertly retrieving a piece of information would affect future memory in the same way as an overt response. In other words, he would disagree with the student above who reads her notes out loud as she studies, "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" (Tulving, 1983, p. 47). Tulving provided no further insight into the issue (nor any data to support his position). His assumption, that any form of access to a memory affects future retention in the same way, raises the question of whether we should trust his intuition as a memory researcher, or the subjective experience of students and teachers. Does reading study notes aloud really help more than reading them silently? Does memory benefit further by writing something down? Or is Tulving right in that the response modalities associated with retrieval have very little effect on the memory itself?

Researchers have long known that retrieval can enhance long-term accessibility (e.g. Gates, 1917). More recently this finding has been labeled the *testing effect* and has become a widely researched topic (for a review, see Roediger & Karpicke, 2006a). One of the most interesting findings is that retrieval practice can often enhance future

retention compared to a restudy control group, even if no corrective feedback is provided, suggesting that it is the act of retrieval itself that enhances learning. Indeed, Tulving's quote above would argue that a covert retrieval, or retrieval without overtly producing the answer via writing or speaking, should still yield a testing effect. It remains unclear, however, if producing an answer, either by saying a word aloud or writing it down, is part of what makes the testing effect occur.

Taking a step back, perhaps we should first ask whether there are any differences in recall between different types of overt response modalities; for instance, can people recall more words from a previously studied list by writing or speaking their answers? We can also ask whether response modality on a first test influences performance on a later test. One possibility is that in writing answers subjects are able to see their responses in front of them, effectively allowing an additional study session. (Of course, this particular concern was stronger when subject responses were collected via paper and pencil, and the subject could scan over their responses multiple times during the recall period.) With this logic, writing or typing responses should yield a stronger testing effect than reporting answers verbally (due to the extra study time) a disconcerting idea considering the majority of testing effect research requires subjects to write or type their responses.

With these ideas in mind, this project seeks to answer several questions related to response modality and retrieval. Will subjects recall more if they respond by speaking rather than writing their responses? Will writing on a first test lead to better recall on a second test compared to reporting answers verbally? Does the testing effect depend on

subjects making overt responses on an earlier test, or will a covert retrieval garner the same benefit? The current research projects answers all of these questions.

Relevant background research is broadly organized into two sections. The first area of research explores how different processing modalities, either at encoding or at retrieval, can influence recall. The second area examines how covert retrieval influences memory, and whether it has the same effects on future recall as overt retrieval.

The Effects of Typing and Speaking on Retrieval

Even with the plethora of research available on the testing effect, the question of how initial response modality affects later retention remains unanswered. Although other research has explored how response modality can influence recognition and recall in memory experiments, to our knowledge it has never been explored within the context of repeated testing. Gardiner, Passmore, Herriot and Klee (1977) did employ a two-test paradigm where subjects wrote some words and spoke others during the first test, but the primary focus was whether subjects could recognize words they had recalled earlier rather than the effects of response modality on later recall. On the final recognition test, the group that both wrote and spoke words during the intermediate test showed better recognition memory for those words than the groups that only wrote or only spoke their answers; there was no difference between the oral and the writing groups. It is important to note that the task on the final recognition test was to identify words they had recalled earlier, rather than words they had originally studied, making the test more of a source monitoring judgment than a pure episodic memory test. Regardless, the implication is that subjects were more accurate at identifying what they had previously recalled if they had produced their responses by writing *and* speaking. Gardiner et al. proposed a model

where retrieval does not just increase the strength of a trace, but also causes qualitative changes in the encoding of the trace. Retrieving an item and producing it orally, for example, results in articulatory and auditory information becoming attributes of the item's memory trace. Likewise, writing a response can cause visual and kinesthetic attributes (such as the visual appearance of the word, or the sensory feedback from writing) to become associated with the item. Responding in different modalities, therefore, can cause different patterns of encoding, depending on the specific response modality. In sum, this view suggests that different response modalities may affect memory in qualitatively different ways. The implication is that various forms of responding may create variable encoding during a first test leading to greater recall on a later test.

Other research has shown that saying a word aloud can enhance retention compared to reading it silently, although these experiments have typically varied modality at initial study, rather than retrieval. Colin MacLeod and his collaborators revived some little-known findings from the seventies and eighties that showed reading aloud can lead to enhanced memory compared to reading silently (Hopkins & Edwards, 1972; Conway & Gathercole, 1987), and named this finding the production effect (e.g. MacLeod, Gopie, Hourihan, Neary, & Ozbuko, 2010). MacLeod et al. (2010) conducted a series of eight experiments exploring the boundary conditions of the effect, concluding that producing (speaking) words creates a distinct verbal record that can facilitate future recognition. Whether production facilitates future recall is unclear, but at least in recognition experiments, it is quite clear that producing a word enhances learning compared to reading silently. If production does function by creating a more distinctive

verbal record, then perhaps the mnemonic benefits of testing may be due in some part to having subjects explicitly report their answers. Although the production effect has so far been limited to an encoding manipulation, it is not unreasonable to assume that similar processes may occur during retrieval and may boost the testing effect, especially as some encoding is thought to occur during retrieval (McDaniel & Masson, 1985).

Covert vs. Overt Responding

Tulving's quote above suggests that a covert retrieval (or retrieving information without producing it) should have similar effects on memory as an overt retrieval would, and indeed research has shown that testing effects can be acquired while having subjects only covertly retrieve information at a first test. Carpenter, Pashler, Wixted, and Vul (2008) had subjects learn obscure facts and later presented questions about the facts, asking subjects to think about the answers without overtly responding. Covert retrieval at a first test led to higher performance on a final test compared to restudying for an equivalent amount of time. Even without comparison to an overt response group, these results suggest that if the testing effect can occur without an overt response on the first test, than the act of retrieval is partly driving the testing effect, although overt responding may increase the effect.

Other research, however, has shown that covert retrievals do not always have the same effects as overt retrievals. Whitten and Bjork (1977) conducted an experiment comparing restudy, retrieval, and covert rehearsal conditions with varying degrees of delay in a hybrid design using elements from the Brown-Peterson and free recall paradigms. Subjects saw a word pair, then performed a digit-shadowing task for a variable amount of time before either recalling the word pair (an overt response), silently

rehearsing the pair (a covert rehearsal), or seeing the word pair again (a re-exposure). After every 12 trials subjects recalled all the words they could remember on a free recall test. The results showed that increasing the delay between the initial presentation and the intermediate test trial resulted in increasing final recall performance for both the restudy and the test conditions, but not for the covert rehearsal condition. In other words, there was no spacing or testing effect obtained for the covert rehearsal condition where subjects were instructed to mentally recall the word. Whitten and Bjork offered two explanations for the lack of improvement in the covert rehearsal condition. The first hypothesis was that "…overt retrieval and covert rehearsal involve qualitatively different processes. Little data exists to support or deny this conclusion" (1977, p. 472). As the rest of their article is concerned with spacing effects, there was no further discussion of whether covert rehearsal and overt responses actually use different processes. The second explanation was a lack of experimental control over what subjects were actually doing during the covert rehearsal conditions; subjects could have rehearsed previous items or done nothing at all.

Here we encounter the obvious difficult in exploring silent or covert tests. As Whitten and Bjork (1977) suggested, the experimenter has no measure of how much subjects are actually retrieving during the covert test, or even if they are performing the required task. Subjects may be actively attempting to retrieve an item, rehearsing items they saw earlier, or planning how they will spend their $10 for participating in the experiment.

Izawa (1976) conducted an experiment directly comparing the effects of overt and covert retrieval on future memory performance using a complex paired-associates design.

Each condition consisted of a study phase followed by five cued-recall tests. This cycle of a study session followed by several tests was repeated five times (using the same materials in each rotation) before a final test. Table 1 shows the different conditions, each with a unique pattern of vocalized (overt) and silent (covert) tests. A vocalized test trial presented the cue and had subjects respond verbally with the target. A silent test trial would also present the cue, but would have subjects bring the target word to mind without overtly producing it. Neither type of test trial included any form of feedback; the only opportunity for subjects to acquire new information or to fix errors was during the study trials at the beginning of each cycle. The baseline condition, for instance, consisted of a study session followed by five vocalized tests (STTTTT). Subjects repeated this cycle five times before taking the final vocalized test. In another condition, subjects again started with a study session, but all of the subsequent tests were silent (Sttttt). Izawa compared six conditions in all, including some that mixed silent and vocalized tests within a cycle (such as STtttt, StttT, and STtttT) and a control condition where subjects sat quietly, without making a response (as a comparison to the silent test conditions). Izawa was interested in two outcomes: how performance changed within a cycle across tests, and how performance on the final test was affected by the different patterns of testing. Izawa's experimental design allowed her both to estimate how silent test trials affected performance within a cycle, and to examine performance over the long term on the final test. The data revealed that vocalized tests prevented short-term forgetting (i.e. forgetting within a cycle) but silent tests did not. In other words the StttT condition showed poorer performance on the last vocalized test than did the STTTTT condition, where every test was vocalized.

Table 1

Design of Izawa (1976). Subjects learn paired associates in one of six conditions. S is a study period, where subjects see all of the word pairs. During T, subjects vocally recall the target word; during t, they think of the target word without responding. B represents a blank trial where subjects sit quietly. Subjects learn the same list of 20 word pairs during each cycle, and after 5 cycles take a final test.

Condition	Cycles 1 – 5	Final Cycle
STTTTT	STTTTT	ST
STtttT	STtttT	ST
STtttt	STtttt	ST
StttttT	StttttT	ST
Stttttt	Stttttt	ST
SBBBBB	SBBBBB	ST

All conditions showed a boost in performance during the next cycle, since each cycle began with a study period. Somewhat surprisingly, however, performance on the final test was equal among all conditions except for the control condition where subjects were asked to rest. Regardless of whether subjects reported items verbally or only silently retrieved them during the early test cycles, final test performance was the same. Izawa interpreted these results as suggesting that although the silent tests might not prevent forgetting within a cycle, they did potentiate future study, meaning subjects would learn more from the next study period (for more on test potentiation see Izawa, 1966). In her conclusion Izawa suggested that either a silent or a vocalized test could potentiate future learning, but that only vocalized tests could prevent short-term forgetting.

Clearly, it is important to consider how subjects respond to covert retrieval instructions. The ideal instructions would always elicit a true covert retrieval, where subjects actually attempt retrieval, regardless of whether they need to report their response. One possible solution is to use a task that requires subjects to retrieve an item in order to complete the task, but does not require them to overtly report the item. Some metacognition tasks may meet this requirement, and indeed researchers have explored how making certain metacognitive judgments can influence future retrieval (Nelson & Dunlosky, 1991; Spellman & Bjork, 1992; Kimball & Metcalfe, 2003). Judgments of learning (JOLs) are a measure of awareness of one's own learning. Subjects are asked a question such as, "how confident are you that you will answer this item correctly on a future test?" and are asked to make a numerical rating on a scale. Although judgments of learning made immediately after studying are typically inaccurate, JOLs made after a

delay are relatively accurate (Nelson & Dunlosky, 1991). The initial explanations for this increased accuracy was that the time delay allowed subjects to evaluate the item in their long term memory without interference from short term memory. A later interpretation, from Spellman and Bjork (1992), suggested that any attempts to evaluate one's own memory would result in a change to the memory being evaluated. In making a JOL subjects attempt to retrieve the target item from memory, and if retrieval is successful will assign the item a high JOL, whereas if it is not successful they will assign the item a low JOL. Thus delayed JOLs may show enhanced accuracy because subjects are covertly retrieving the item without overtly producing it. Kimball and Metcalfe (2003) reported further research supporting this stance: "The results indicate that the delayed-JOL effect stems from differences in spaced study opportunities for high-JOL and low-JOL items in that condition, caused by differences in the success of covert retrieval for those items at the time of the delayed JOL," (p. 926). Although other attributes may influence how a subject makes a JOL, such as retrieval fluency or other elements of the cue (Dunlosky & Metcalfe, 2009, pp. 104-110) there is strong evidence to suggest that at least with a cue-only delayed JOL, the success or failure of a covert retrieval is an important factor in shaping JOLs. As such, a delayed, cue-only JOL likely requires a covert retrieval, and thus may be more efficient in eliciting covert retrievals compared to generic instructions that only ask subjects to "think about the answer."

Present Research

The current set of experiments aims to dissociate the effects of memory retrieval from any byproducts of production, as well as examining any differences in recall between typing and speaking. All of the experiments used a paired associates procedure,

consisting of a study, intermediate, and final test phase. Response modality was varied at the first or final tests (or both). Experiment 1 compared how different response modalities, such as typing, speaking, and making a JOL, influence future test performance. Experiment 2 was a replication with two procedural changes that ultimately yielded a much stronger testing effect. Experiment 3 introduced a new procedure with timing deadlines and different response options to allow a more direct comparison between the effects of covert and overt retrieval on later retention.

Experiment 1

Experiment 1 examined how different response modalities during a first test influenced performance on a final cued recall test within a paired-associates paradigm. The experiment had three phases: an initial study phase, an intermediate phase in which response modality was manipulated, and a final test where response modality was also manipulated. During the intermediate phase subjects recalled words by speaking, typing, or through a covert retrieval. In two control conditions subjects either restudied the information (the *restudy* control condition) or had no further exposure to it (the *study once* control condition). After each retrieval or restudy trial, subjects made a JOL for the current item, predicting their performance on the final test. The *covert* condition had subjects make the JOL without requiring them to produce the item via typing or speaking. During the final test (after a two-day delay) subjects responded by either typing or reporting their answers verbally.

Three distinct predictions can be made about performance on the final test. First, as Tulving hypothesized, any form of retrieval may have similar effects on memory; thus he would predict that the *aloud*, *type*, and *covert* retrieval conditions should all yield

similar performance on the final test. As those conditions are all different forms of retrieval practice, they should lead to better performance than the control conditions. Another hypothesis, influenced by research on the production effect, is that the overt response conditions, *aloud* and *type,* may facilitate recall more so than the *covert* condition. Finally, one could hypothesize that having a match between response modality at the first test and response modality at the final test (e.g. typing on both the first and final test) would maximize performance. This prediction follows from the transfer appropriate processing principle, which states that congruency between processes at encoding and processes at retrieval will maximize performance at retrieval (Morris, Bransford, Franks, 1977).

Method

Subjects

Fifty subjects from Washington University in St. Louis' research pool participated for course credit or payment. Paid subjects earned $10 for their time. Five subjects had incomplete data either due to a computer programming error or for failing to return to the lab for the second session and were consequently replaced with five new subjects.

Stimuli

Seventy-five weakly related word pairs were generated from The University of South Florida word association, rhyme, and word fragment norms (Nelson, McEvoy, & Schreiber, 1998). The selected words pairs had a forward cue-to-target strength and backward target-to-cue strength between .01 and .02 and had between three and nine letters per word. Word pairs were also within a median range of concreteness and

frequency and were all nouns. Example pairs are: "airplane – trip" and "blossom – cherry." No duplicate words appeared within the lists.

Design and Counterbalancing

The experiment was a 5 (intermediate response modality: *type, aloud, covert, restudy, study once*) x 2 (final test response modality: *aloud* vs. *type*) mixed design. Intermediate response modality was manipulated within subjects while final test modality was manipulated between subjects. The 75 word pairs were randomly divided into five groups of 15 words each and were rotated through the five conditions during the intermediate phase (*type, aloud, covert, restudy,* and *study once*). The different conditions were blocked together for presentation, and presentation order was counterbalanced between subjects. The final test manipulated response modality between subjects, with one group responding verbally, and the other half responding by typing.

Apparatus

Stimuli and responses were presented and recorded on a PC using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). During the oral test phases, an experimenter recorded subject responses.

Procedure

The experimental procedure consisted of a study phase, an intermediate phase where subjects were tested on some words and restudied others, and a final test phase. Subjects were tested individually, and upon arrival were informed about the nature of the experiment and gave their informed consent. They were told they would be learning word pairs and would take one test today and another when they returned to the lab in two days.

After reading the experimental instructions, subjects entered the study phase. Word pairs were presented in black, lowercase letters on a white background for 4 s, followed by a 500 ms inter-stimulus-interval. All 75 word pairs were presented in random order, and the computer cycled through the entire list twice. The entire study phase lasted around ten minutes.

Subjects then immediately moved into the intermediate phase. Four of the conditions were presented in a blocked order, while the items in the *study once* condition were not presented. Before the start of each block, instructions appeared informing subjects as to how they should respond (e.g. "For this block please say the target word aloud after seeing the cue word").

In the *type* condition, subjects saw the cue word and a set of question marks (e.g. airplane - ?????), and had six seconds to type the target word into a box on screen. After six seconds the screen changed to display a JOL prompt: "How confident are you that you will correctly recall this word pair at the final test two days from now?" Subjects responded using the keyboard and made their rating on a scale from zero (no chance of recalling) to one hundred (absolutely sure I will recall it). There was a 500 ms inter-stimulus-interval (a blank screen) before the next trial. Likewise, the *aloud* condition was identical except that subjects responded orally, rather than typing their answers. An experimenter was present in the room to record their responses. In the *covert* condition the cue was presented with questions marks, as in the *type* and *aloud* blocks, but no additional processing instructions were displayed. After six seconds subjects made their JOL and moved on to the next trial. In the *restudy* condition subjects saw both the cue and the target word during the six-second window before making a JOL. The 15 words in

the *study once* condition were not presented. After completing all blocks, subjects were dismissed with instructions to return to the lab in two days.

When subjects returned to the lab they took a final cued recall test on all of the word pairs. A cue word was presented on screen and subjects were asked to generate the target word; they had as much time as they desired. Half of the subjects responded by typing their answers on screen, where their response was displayed until pressing enter, after which the next trial began. The other half of subjects responded orally. After saying their response, the experimenter recorded their answer and pressed a key to move onto the next trial. After finishing the cued recall test subjects were thanked and debriefed.

Results

Answers were coded as correct if it was an obvious misspelling of a target word (e.g. “blossum” instead of “blossom”).

First Test Performance

Measures of recall are only available for the *type* and *aloud* conditions on the first test as the other three conditions did not allow overt responses. As seen in the first column of Table 2, the *aloud* and *type* conditions were not significantly different from one another, indicating that response modality did not influence performance on the first test, $t(49) = .93$, $p = .359$. It is important to note that subjects failed to recall over 30% of the items in the overt testing conditions, a fact that will be important in interpreting the final test results.

Table 2

Test 1 performance, JOL ratings and resolution in Experiment 1. The first column shows percent recalled during the first test. The second column shows mean JOLs (made during the intermediate phase) for each condition; subjects predicted performance on the final test on a scale from 0 - 100. The third column is resolution, a relative measure of JOL accuracy in predicting final test performance.

	Test 1 Recall	JOL	Gamma
Type	0.68 (.04)	54.3 (3.0)	0.73 (.03)
Aloud	0.66 (.04)	54.9 (3.1)	0.80 (.06)
Covert	--	57.9 (3.0)	0.66 (.06)
Restudy	--	61.7 (2.4)	0.32 (.06)

Note. Resolution is calculated via the Goodman-Kruskal gamma correlation. Standard errors are displayed in parentheses.

Judgments of Learning

Subjects made JOLs after each trial, where they predicted their performance on the final test. There are several ways to evaluate JOL data, including both absolute and relative measures. Both of these will be discussed below.

JOLs. Table 2 shows the average JOLs for each condition. A one-way within subjects ANOVA revealed a significant main effect, $F(3,147) = 5.816$, $p = .001$. Planned comparisons revealed that the *type* and *aloud* conditions yielded significantly lower JOLS than the restudy condition, $t(49) = 3.21$, $p = .002$ and $t(49) = 3.35$, $p = .002$ respectively, suggesting that subjects tended to be more confident after restudying than after attempting retrieval. This finding is consistent with other research that shows that subjects assign high JOLs to restudied items, possibly due to increased fluency as a result of restudying (Karpicke, 2009). Further, although there were some other differences between the mean JOLs in the different conditions (with *aloud* being lower than *covert*, and *covert* being lower than *restudy*), those differences did not reach significance (*covert* vs. *restudy*, $t(49) = -1.80$, $p = .077$; *aloud* vs. *covert,* $t(49) = -1.79$, $p = .080$).

Absolute JOL Accuracy (Calibration). Figure 1 shows calibration curves, with JOL ratings on the ordinate and final test performance on the abscissa, with the straight diagonal line indicating perfect accuracy. The JOL rankings were split into 7 bins (0, 1-20, 11-40…100), and the final test results were averaged across each bin. Data points above the diagonal line indicate underconfidence, while data points below the diagonal line indicate overconfidence. Overall, subjects tended to be fairly accurate, with a slight indication of underconfidence for the *aloud* and *type* conditions. These results imply that in general, subjects are fairly well calibrated in predicting their overall performance.

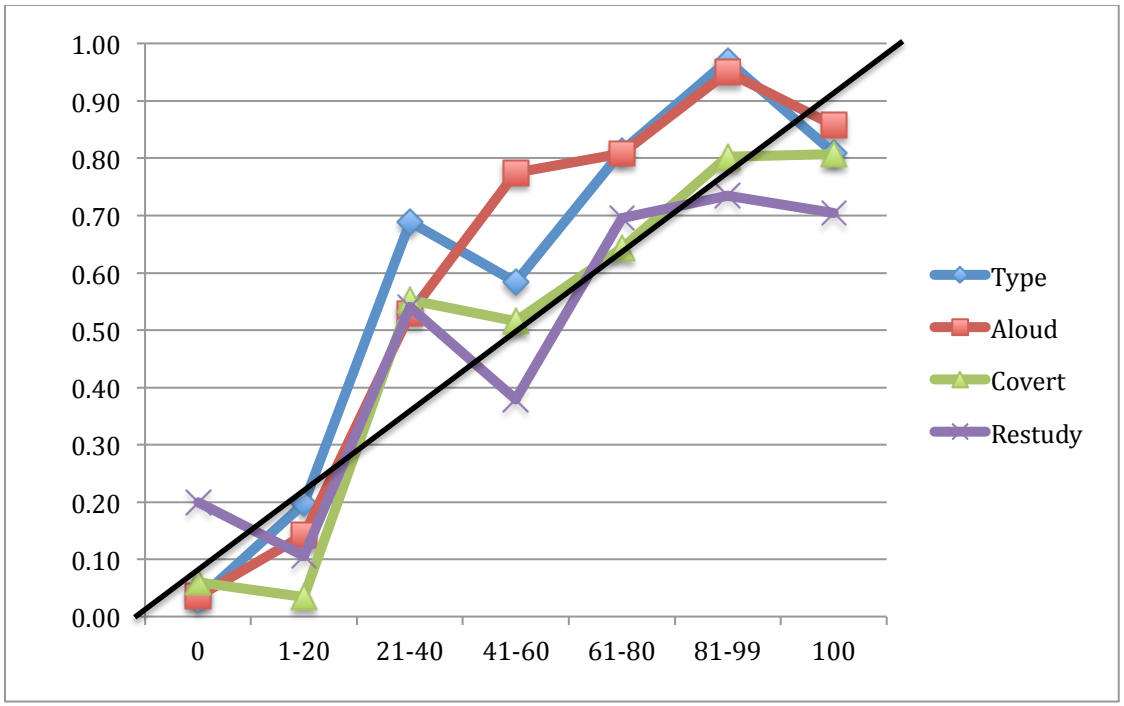

Figure 1. Displays calibration curves, representing subjects' accuracy in predicting their performance on the final test. The curves represent absolute correspondence between actual test performance on the final test and their estimated performance.

Relative JOL Accuracy (Resolution). The third column of Table 2 presents the gamma scores for the different conditions, a measure of resolution, with 1.00 representing a perfect match between estimated and actual performance on the final test. A 4 (intermediate phase response modality: *type*, *aloud*, *covert*, *restudy*) X 2 (final test response modality: aloud vs. typed) mixed model ANOVA revealed a significant main effect of first test response modality on resolution, $F(3,111) = 18.42$, $p < .001$. The *aloud* condition led to significantly more accurate predictions than in the *covert*, $t(39) = 2.22$, $p = .033$, and *restudy*, $t(41) = 7.04$, $p < .01$, conditions, but was not significantly different from the *type* condition. The *type* condition had significantly higher resolution than the *restudy* condition, $t(43) = 4.57$, $p < .001$. Finally, the *covert* condition was significantly higher than the *restudy* condition, $t(41) = 4.29$, $p < .001$. Attempting retrieval before making a JOL led to more accurate JOLs, meaning subjects were better able to distinguish items they would eventually answer correctly from items they would eventually miss, compared to only making the JOL or restudying.

Final Test Performance

Table 3 shows the proportion of words correctly recalled on the final cued recall test. A 5 X 2 repeated measures ANOVA with response modality during first test as a within-subject variable and response modality at final test as a between-subjects variable revealed a significant main effect of response modality at the first test, $F(4,192) = 25.98$, $p < .001$. All four re-exposure conditions during the intermediate phase led to better performance on final test performance relative to the *study once* condition. However, there were no further differences between the groups as the *type, aloud, covert,* and *restudy* conditions all led to similar performance on the final test. There was a marginally

significant main effect of final test modality ($F(1,48) = 3.84$, $p = .056$), with the oral group appearing to be better than the typed group; however, this difference between groups did not reach significance, despite the oral group showing stronger recall across all conditions compared to the typed group. (This effect was not replicated in later experiments). There was no significant interaction ($F(4,192) = .245$, $p = .913$). Paired t-tests revealed that words only presented during the initial study period (*study once)* were recalled significantly less well during the final test than words in any other conditions. In other words, the *type, aloud, covert,* and *restudy* conditions all yielded higher performances than the *study once* condition. The t-test scores comparing the different conditions to the *study once* condition were as follows: *type,* $t(49) = 12.42$, $p < .001$; *aloud,* $t(49) = 11.83$, $p < .001$; *covert,* $t(49) = 10.79$, $p < .001$; and *restudy* $t(49) = 5.60$, $p < .001$. There were no other significant differences among those four conditions.

Table 3 also shows the performance on final test broken down by response modality during the final test. As stated above, the repeated measures ANOVA did not reveal a main effect of response modality during the final test on final test performance, though there was a trend towards the oral group having better performance. T-tests comparing performance between groups within each condition did not reveal any significant differences.

Table 3

Experiment 1. Mean recall on final test after experiencing different response modalities during the first test, relative to response modality at final test.

	Type	Aloud	Covert	Restudy	Study Once
Final test modality					
Type	.51 (.05)	.52 (.06)	.49 (.05)	.51 (.05)	.24 (.05)
Oral	.65 (.06)	.63 (.06)	.62 (.05)	.60 (.05)	.33 (.05)
Both Groups	.58 (.05)	.58 (.06)	.56 (.06)	.55 (.05)	.29 (.05)

Note. Standard errors of the mean are in parentheses.

Discussion

Overall, the results of Experiment 1 suggested that response modality had very little, if any effect, on retrieval. First, there were no differences between the amount subjects were able to recall either via typing or speaking on the first or final test. Although at the final test the oral group appeared to recall more, that difference was not significant. Second, the specific response modality on a first test did not seem to influence the probability of recall on the final test, as performance was equivalent across all conditions (except for *study once*). Finally, covert retrievals appeared to be just as effective as responding overtly. Despite having no data, when Tulving claimed that all forms of retrieval would have a similar impact on later retention, he appears to have been correct. Our results showed no differences between responding orally or by typing, and even a covert retrieval, elicited by having subjects make a JOL, induced remarkably similar performances to the two overt response conditions.

One troublesome finding, however, is that the *restudy* condition showed equal performance to the retrieval conditions (*aloud, type,* and *covert)* on the final test. Previous research (Roediger & Karpicke, 2006b) suggests that retrieval practice during the intermediate phase should lead to higher performance on the final test compared to the restudy condition. On the one hand, if we define a testing effect as showing enhanced recall for retrieval practice conditions compared to a restudy condition, then we did not find a testing effect. On the other hand, if we take the baseline condition to be the *study once* condition, then we can claim a testing effect, since having subjects recall some words led to better performance on the final test.

Thus by one comparison we found a testing effect, but by another we did not. As mentioned above, previous research has consistently shown retrieval practice to lead to better recall than restudy conditions; why did our results not show this? A surprise in our data relative to other similar experiments is that the *restudy* condition produced nearly 100% improvement in recall (.55 vs. .29). In most experiments, a single restudy session does not yield such large dividends. One explanation is that having subjects make a JOL during the *restudy* condition caused a covert retrieval or some other sort of deep processing, explaining why performance on the *restudy* condition was equated with the other retrieval conditions. Perhaps the JOL caused the unusually high benefit from the additional study session.

Additionally, by not providing feedback during the testing conditions, subjects were only re-exposed to the word pairs they could correctly recall. Since subjects only correctly answered on average .68 for the *type* condition and .66 for the *oral* condition, they were not re-exposed to the remaining word pairs. We assume that in the covert retrieval condition the figure was the same, but of course we cannot know. During the restudy block, however, subjects were re-exposed to all of the word pairs in that condition. We could expect the testing conditions to result in higher performance had initial recall during the first test been higher (for a detailed analysis of this issue see Wenger, Thompson & Bartling, 1980). Experiment 2 will address both of these issues through two procedural changes.

Although we did not find a strong testing effect, there was a marginally significant difference between the oral and typed group at the final test. There are three possible explanations: One, subjects in the oral group were simply more able subjects;

two, having an experimenter in the room during the oral test may have caused subjects to be more vigilant; and three, perhaps subjects can actually recall more verbally. We intentionally used lenient scoring criteria for the typed responses, as strictly coding would have eliminated many items that were misspelled in that group. Obviously subjects will not misspell items when they are responding verbally.

Although not the main focus of the project, there are some interesting results with the JOL data, with the overt retrieval conditions, *type* and *aloud,* typically yielding JOLs that were lower overall in terms of the average scores, yet ultimately more accurate than the *covert* and *restudy* conditions as measured by resolution (see Table 2). Additionally, the *covert* condition tended to be relatively more accurate than the *restudy* condition. The fact that the overt retrieval conditions are more accurate than the *restudy* condition is not surprising; indeed this is the delayed JOL effect (Nelson & Dunlosky, 1991). In the *restudy* condition subjects are exposed to both the cue and target before they make their JOL; previous research has shown that when the target is also displayed the higher accuracy associated with the delayed JOL effect disappears (Dunlosky & Nelson, 1992). Somewhat more interesting is the fact that the judgments for items in the *aloud* condition were more accurate than for those in the *covert* condition (as measured by gamma), given that many researchers have hypothesized that subjects make delayed JOLs by covertly retrieving the target item, and using the success of failure of that covert retrieval to inform the JOL (Spellman & Bjork, 1992; Kimball & Metcalfe, 2003). As the results of Experiments 2 and 3 will show, making a JOL may not be a valid substitute for a covert retrieval. To summarize the metacognitive results, making an overt retrieval attempt led

to more accurate JOLs than just making the JOL without attempting an overt retrieval beforehand.

Experiment 2

Experiment 2 introduced two procedural changes in hopes of obtaining a stronger testing effect. The first change was eliminating the additional JOL procedure after presenting the cue in the *type, aloud,* and *restudy* conditions. As mentioned earlier, having subjects make a JOL after restudying may cause some additional processing that benefits future retrieval. Although previous research (Nelson & Dunlosky, 1991; Kimball & Metcalfe, 2003) suggests that an immediate JOL (in this case making a judgment after seeing both the cue and the target) does not provide an additional benefit to retrieval, one interpretation of the results of Experiment 1 is that the *restudy* items were enhanced by making a JOL after restudying. However, Experiment 2 still included a *covert* condition where subjects were provided with a cue and subsequently made a JOL.

The second change was to provide feedback during the *type, aloud,* and *covert* conditions. Providing feedback allowed participants to correct any mistakes made on the first test and re-exposed all of the word pairs. Feedback has been shown to be an effective way to further increase any benefits resulting from retrieval practice with a delayed final test (Butler & Roediger, 2008) Thus, the *type, aloud,* and *covert* conditions should show a boost in final test performance relative to the restudy condition.

In summary, in Experiment 2 subjects studied paired associates. During the intermediate phase they recalled some words by typing or speaking, made a JOL for some words, and restudied others (with one group of words not being exposed during this phase). Total exposure time for each pair was equated. After a two-day delay, subjects

returned to the lab and took a final cued recall test that was varied between subjects (oral or typed). Despite being a close replication of Experiment 1, eliminating the JOL after each word presentation and providing feedback should enhance performance on the final test for the *type, aloud,* and *covert* conditions relative to the *restudy* condition. Our experimental hypotheses remain the same.

Methods

Subjects

Fifty subjects from the same pool as Experiment 1 participated for course credit or cash. Six subjects failed to return to the lab for the second day of testing and were consequently replaced with six new subjects.

Materials

The same materials from Experiment 1 were used, seventy-five weakly related word pairs.

Design and Counterbalancing

As before, the experiment was a 5 (intermediate response modality: *type, aloud, covert, restudy, study once*) x 2 (final test response modality: *aloud* vs. *type*) mixed design. First test response modality was manipulated within subjects while final test modality was manipulated between subjects. The 75 word pairs were randomly divided into five groups of 15. The five groups of word pairs were rotated through the five conditions in the intermediate phase (*type, aloud, covert,* and *restudy)* or were not presented. The final test was between subjects, being either entirely typed or entirely oral.

Procedure

The experimental procedure was very similar to Experiment 1. During the study phase, subjects were presented with the weakly related word pairs and saw the entire list twice. Subjects then moved into the intermediate phase where they were tested on some words and restudied others. As before, the different conditions were blocked. In the *type* and *aloud* blocks subjects were presented with the cue word and had five seconds to respond by appropriately typing or saying the target word. The correct answer was then displayed for two seconds, and subjects moved on to the next trial. Note that subjects did not make a JOL after attempting retrieval of the target word and that correct answer feedback was provided. In the *covert* block subjects saw the cue word and made a JOL. They had five seconds to respond and the correct response was displayed for two seconds after making the JOL. In the *restudy* condition subjects saw both the cue and the target word for seven seconds, equating the total exposure time per item with the retrieval conditions. Of course subjects saw the target item for the entire seven seconds in the *restudy* condition compared to only two seconds of exposure in the retrieval conditions, which should favor the *restudy* condition (also note the lack of a JOL after restudying). Finally, one group of 15 words was not presented during the intermediate phase, the *study once* control group. Subjects left, then returned two-days later to take a cued recall test on all of the word pairs. Half of the subjects responded orally, and the other half responded by typing their answers.

Results

As in Experiment 1, the outcomes of interest for Experiment 2 were performance on the first test for the *aloud* and *type* conditions, and performance on the final test, both as a factor of response modality at the first test and response modality at the final test. To

anticipate the results, the *aloud* and *type* conditions led to increased performance on the final test compared to the *covert* condition, which in turn outperformed the *restudy* condition. The *study once* condition led to the worst performance on the final test.

First Test Performance

During the first test, recall data were only available for the *aloud* and the *type* conditions. Unexpectedly, subjects performed better in the *type* condition (M = .70, SD = .25), than the *aloud* condition (M = .60, SD = .31). A paired samples t-test, $t(49) = 2.78$, $p = .008$, revealed the difference was significant. Because subjects received feedback on all trials, exposure differences were minimized.

Final test performance

Table 4 shows the proportion of words recalled on the final cued recall tests, broken down by response modality at the final test and response modality during the first test. A 5 x 2 repeated measures ANOVA with intermediate phase response modality as a within-subjects variable and final test response modality as a between subjects variable revealed a main effect of intermediate phase response modality, $F(4,192) = 80.05$, $p < .001$. However, there was no main effect of response modality at the final test, nor was the interaction significant. Since there was no effect of final test response modality on performance, those two groups were collapsed for the remaining analyses.

Planned comparisons revealed several effects of intermediate phase response modality on final test performance. First, there was no significant difference between the *type* and *aloud* conditions, $t(49) = .40$, $p = .69$, although both of those conditions yielded significantly higher recall than the *covert* condition, (type vs. covert: $t(49) = 3.91$, $p < .001$; aloud vs. covert: $t(49) = 4.39$, $p < .001$), the *restudy* condition (type vs. restudy:

Table 4

Experiment 2. Mean recall on final test after experiencing different response modalities on the first test, relative to response modality at final test.

	Type	Aloud	Covert	Restudy	Study Once
Final test modality					
Type	.69 (.06)	.70 (.05)	.62 (.05)	.49 (.05)	.37 (.06)
Oral	.66 (.05)	.67 (.04)	.56 (.05)	.37 (.04)	.26 (.03)
Both Groups	.68 (.04)	.69 (.03)	.59 (.03)	.43 (.03)	.31 (.03)

Note. Standard errors of the mean are in parentheses.

$t(49) = 7.99$, $p < .001$; aloud vs. restudy: $t(49) = 8.54$, $p < .001$), and the *study once* condition (type vs. study once: $t(49) = 12.19$, $p < .001$; aloud vs. study once: $t(49) = 15.38$, $p < .001$). Second, the *covert* condition yielded significantly higher recall than the *restudy* and *study once* conditions (covert vs. restudy: $t(49) = 5.52$, $p < .001$; covert vs. study once: $t(49) = 11.88$, $p < .001$). Finally, the *restudy* condition yielded higher recall than the *study once* condition (restudy vs. study once: $t(49) = 4.97$, $p < .001$). To summarize, the *aloud* and *type* conditions yielded the best recall on the final test, followed by the *covert* condition, then the *restudy* condition, and finally the worst recall was yielded by the *study once* condition.

Discussion

The results of Experiment 2 showed a stronger testing effect compared to Experiment 1, in that the test conditions led to higher performance than in the restudy condition. The conditions where subjects were required to make an overt response (*aloud* and *type)* yielded the best performance on the final test. Although a testing effect was also obtained in the *covert* condition, the benefit to retrieval on the final test was not as strong. This outcome suggest that testing effects may be driven both by the act of retrieval and by the production of an answer. In other words, it appears that covert retrieval may not be as potent a memory enhancer as an overt retrieval.

Why was the pattern of results different from Experiment 1? Providing feedback should have improved all of the retrieval conditions (i.e. *aloud, type*, and *covert*), as subjects had the chance to correct any mistakes they made on the first test (errors of omission or commission). Dropping the additional JOL procedures also led to a more streamlined design and may have affected the relative performance of the *restudy*

conditions. Because exposure was equated and feedback should affect all of the retrieval conditions equally, the difference in outcome between experiments is probably due to the absence of a JOL in Experiment 2. As mentioned earlier, subjects showed nearly a 100% improvement from the study once condition to the restudy condition in Experiment 1, but had a much smaller improvement in Experiment 2 (.31 vs. .43) Perhaps subjects engaged more fully in retrieval when they had to make a JOL in Experiment 1.

Regardless, the finding that the overt response conditions yielded better recall than the covert retrieval condition is consistent with some earlier research: the production effect suggests that producing a word should yield benefits over silent retrieval, and Izawa (1976) suggested there are some differences between covert and over retrieval. However, the finding that performance in the two overt retrieval conditions was better than in the covert retrieval condition is a departure from the results from Experiment 1 and is inconsistent with other findings. Although Izawa (1976) did find some subtle differences between covert and overt retrieval, ultimately she concluded that both forms of retrieval contributed equally to future long-term recall. Carpenter et al. (2008) also showed that testing effects could be acquired with covert retrieval practice. Why should the covert condition not be as effective? One possibility is that the JOL procedure in Experiments 1 and 2 might not be a perfect substitute for a covert retrieval. Subjects were presented with the cue word without any additional instructions for a short period of time (whereas in *type* condition subjects would be typing their answers), before making a JOL. Although many researchers have suggested that in this situation subjects would make their JOLs by covertly retrieving the item, subjects may be making those judgments based on other factors, such as retrieval fluency or other aspects of the cue (Dunlosky &

Metcalfe, 2009, pp. 104-110). If this is the case then subjects may not always be covertly retrieving the item, hence the slightly worse performance compared to the overt retrieval conditions. Experiment 3 utilized a new procedure to elicit covert retrievals, by asking subjects to retrieve the item before knowing whether or not they will need to produce it, allowing a better comparison between covert and overt retrieval. More specifically, subjects were presented with a cue word and asked to retrieve the item, but not to report anything until given a cue after four seconds had passed. Then subjects were asked to either report the item or to just report whether or not they could remember it. This procedure essentially forced subjects to retrieve the item, even if they might not have to report it.

One unexpected finding, given the results of Experiment 1, is that the *type* condition led to better performance than the *aloud* condition on the first test, yet on the final test, those two conditions yielded similar levels of recall. In other words, items in the *aloud* condition improved from the first test to the second test (helped in part by corrective feedback), whereas those in the *type* condition showed no improvement (despite their having been provided with feedback). Experiment 1 had two contrary findings: first, performance on the *type* and *aloud* first test was equal, and second, performance on the second test was slightly better in the *oral* condition. Not only were there differences in how much subjects were able to recall from one experiment to the next, the results of Experiment 1 suggested a slight advantage for the oral group at the final test (though not significant) whereas Experiment 2 suggested a slight advantage for the type condition at the first test. These inconsistencies between experiments leaves ambiguity about whether subjects can recall more when they are typing or when they are

speaking. A further question is why the *aloud* condition would show an improvement with feedback, while the *type* condition would not. One possibility is that both overt forms of testing can enhance future retrieval in different ways: saying a response aloud can enhance recall through the production effect, while typing answers allows to see the response on screen for a longer period of time. Or perhaps the better performance of the *type* condition is just an anomalous finding.

Despite one unusual finding, Experiment 2 did show that response modality at the final test did not affect performance. Additionally, it appears the overt retrievals may boost the testing effect relative to a covert retrieval, although this outcome is inconsistent with that of Experiment 1. Experiment 3 utilizes a new way to obtain a covert retrieval, thus avoiding any complications that may arise from using JOLs as a substitute for covert retrieval.

Experiment 3

Experiment 3 is designed to further investigate the relationship between covert and overt retrieval and their influence on later tests. The results of Experiment 2 suggested that an overt retrieval enhances performance on a later test compared to a covert retrieval. There may be issues, however, with using a JOL as a form of covert retrieval. The major difficulty in covert responding is an inability to determine whether subjects are actually retrieving as they are instructed to do. To help address this issue, Experiment 3 implemented a procedure wherein subjects were asked to retrieve a target item, but did not know whether or not they needed to report the item until several seconds had passed. More specifically, subjects were presented with a cue word, but were not able to respond for four seconds. After four seconds they were either cued with "recall"

and were asked to report the target word or with "do you remember?" and were asked to report whether or not they remembered the target word. Subjects had only 1.5 seconds to respond. The timing procedure effectively forced subjects to recall the word initially (if possible) without knowing whether or not they needed to report the word until after completing the retrieval. Thus we have a potentially more reliable method of eliciting a covert retrieval.

Method

Subjects

25 subjects from Washington University in St. Louis's subject pool participated for cash or course credit. Six subjects failed to return for the second day of the experiment and were subsequently replaced.

Stimuli

The same materials from Experiments 1 and 2 were used, weakly related word pairs.

Design

The experiment used one independent variable, the type of reporting activity during an intermediate phase, manipulated within subjects. During the intermediate phase subjects retrieved some items overtly (the *overt* condition) by reporting the target word aloud, retrieved other items covertly (the *covert* condition) by reporting whether or not they could remember the target word, and restudied other items (*restudy)*. One group of items was not presented during this phase, the *study once* control group. Unlike in Experiments 1 and 2 the final test was typed for all subjects.

Procedure

As before, the experiment consisted of a study phase, an intermediate phase, and after a two-day delay, a final cued recall test. Additionally, subjects first completed a short practice phase allowing them to understand the procedure and to ask any questions. After finishing the practice phase they moved onto the experiment proper.

During the study phase 64 word pairs were presented for three seconds each in a random order. Subjects played a video game for two minutes before entering the intermediate phase. During this phase subjects recalled 16 pairs overtly (the *overt* recall condition), 16 pairs covertly (the *covert* recall condition), and restudied 16 others (the *restudy* condition). 16 word pairs were not presented during this phase, the *study once* control condition. Unlike the previous experiments where the different conditions were blocked together, the different trial types were mixed together in Experiment 3 to prevent subjects from knowing until the last moment whether they would actually report the target item.

In the *restudy* condition, the cue and target words both appeared ("airplane – trip") and were displayed for 7.5 seconds. At the beginning of the trials for the retrieval conditions, however, the cue word appeared, but the target word was replaced with a series of question marks (airplane - ?????) . Subjects were instructed to bring the target word to mind, if possible, during a four-second period, but were not to make any indication that they had or had not succeeded in doing so. After four seconds, one of two events occurred. In the *overt* retrieval trials, the word "Recall!" appeared on screen, and subjects reported the target word aloud, or said nothing if they could not remember the word. In the *covert* retrieval trials, the prompt "Did you remember?" appeared, and subjects responded with "yes" if they remembered the word and "no" if they did not.

After being primed with the appropriate cue, subjects had 1.5 seconds to respond. In both cases an experimenter recorded their responses, either true accuracy for the *overt* trials or subject-reported retrieval success for the *covert* trials. After making their response, a feedback screen displayed the correct answer for two seconds (thus equating total exposure time per item for the retrieval conditions with the *restudy* condition). A one second inter-stimulus-interval indicated the start of the next trial.

After finishing the intermediate phase subjects were dismissed and returned to the lab two days later to take a final cued recall test on all of the word pairs. The final test was entirely typed and had no time limit. As the previous experiments did not suggest any differences between typing and speaking, we thought it was adequate to only have a typed final test.

Results

As before, the major outcome of interest is final test performance as a function of first test response mode. Answers were coded as correct if they were a clear mis-spelling of the target world (i.e. “blossum” instead of blossom”).

First Test Performance

On the first test, subjects recalled .42 (SD = .22) of the word pairs correctly in the overt retrieval condition. Comparatively, in the covert condition, subjects self-reported remembering .51 (SD = .20) of the word pairs. As one condition is an overt retrieval, while the other is self-reported retrieval, we must use caution in making comparisons. However, the difference between actual performance and self-reported performance on the first test does suggest that subjects were overconfident in the covert retrieval

condition. A paired samples t-test, $t(24) = -2.28$, $p = .03$, revealed that this difference was significant.

In the *covert* condition, subjects were reporting whether they thought they could accurately recall the target item. These subjects were extremely well calibrated, self reporting to remember 51% of the items at the first test, and actually recalling 51% at the final test. On the final test subjects recalled 25% of the items they claimed to not have remembered at the first test, and 74% of the items that they claimed to remember.

Final Test Performance

Figure 2 shows proportion of items recalled on the final test. A one-way (first phase processing: *overt, covert, restudy,* or *study* once) repeated measures ANOVA revealed a significant effect, $F(3,72) = 46.23$, $p<.001$. Planned pairwise comparisons showed that both the *overt* and c*overt* retrieval conditions led to better performance on the final test compared to the *restudy* condition, $t(24) = 5.72$, $p<.001$ and $t(24) = 4.86$, $p<.001$, and compared to the *study once* condition, $t(24) = 9.69$, $p<.001$ and $t(24) = 8.72$, $p<.001$, respectively. In other words a testing

effect was found, regardless of whether our baseline comparison is the *study once* condition or the *restudy* condition. Further, the *restudy* group showed increased performance on the final test compared to the *study once* group, $t(24) = 5.33$, $p<.001$. Finally, and most importantly, there was no significant difference between the *overt* and *covert* retrieval conditions, $t(24) = .411$, $p= .685$.

Discussion

The critical question of Experiment 3 was to determine whether making a covert retrieval had the same effect on future recall as retrieving and producing an item under

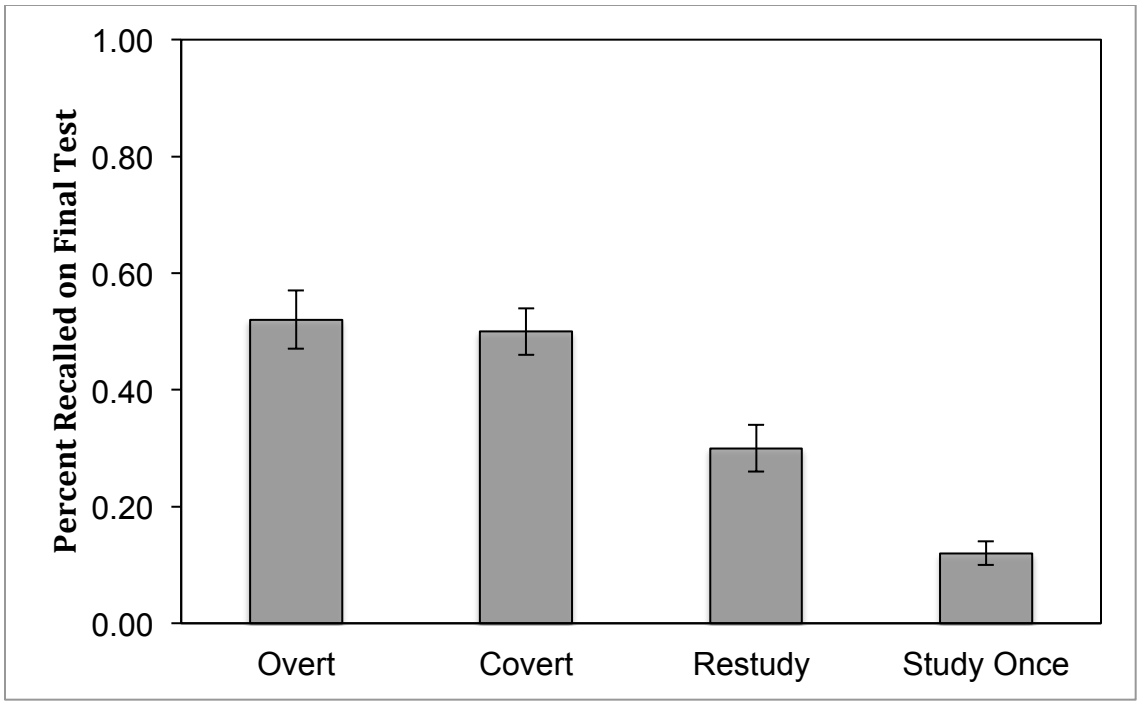

Figure 2. Proportion of items recalled on the final test as a function of processing condition during the intermediate test phase for Experiment 3. Error bars represent the standard error of the mean.

conditions designed to encourage covert retrieval. One of the disadvantages to Experiments 1 and 2 is that a JOL was used as a placeholder for a covert retrieval. Although some research does strongly suggest that delayed JOLs are based on a form of covert retrieval (Whitten & Bjork, 1977; Kimball & Metcalfe, 2003), there is no objective measure of what subjects actual do as they complete a JOL, or whether they retrieve on every trial. The testing procedure in Experiment 3, however, forced subjects to bring items to mind before they knew whether or not they would need to report the answer, thus creating a scenario where subjects were more likely to covertly retrieve information.

During the intermediate phase, subjects claimed to remember more words in the *covert* retrieval condition compared to words actually recalled in the *overt* retrieval condition. Although this comparison must be interpreted cautiously, it does suggest that subjects were overconfident in their ability to remember, which is not surprising, considering subjects are often overconfident in their metamemory judgments (Dunlosky & Nelson, 1994). On the final test, however, words from the *covert* and *overt* condition had similar levels of recall: both were higher than the *restudy* and *study once* control conditions, indicating similar testing effects for words in both conditions. This result suggests that covertly retrieving an item on a first test has the same effect on future retrieval as does making an overt response.

This finding is consistent with the results of Experiment 1, along with several other experiments that have investigated silent or covert retrieval (Izawa, 1976; Smith, 2011), but is inconsistent with the results of Experiment 2, where making a JOL led to worse performance on the final test than an overt response. In summary, the results of

Experiment 3 suggest that an overt production of the answer is not necessary to elicit a testing effect.

General Discussion

The purpose of these experiments was to investigate how different response modalities influenced retrieval both on initial and delayed tests. More specifically, we examined whether subjects can recall more via writing or speaking, whether writing (or speaking) on a first test can lead to better performance on a second test (and whether the type of second test would matter), and whether any form of overt retrieval on a first test leads to better performance on a final test compared to just thinking about a response. All of these questions were aimed at determining whether the beneficial effects of testing arise from retrieval processes or are related to production.

Summary of Results and Qualifications

In Experiment 1, the results did not show a testing effect, at least by one definition of the testing effect. Although the *type*, *aloud*, and *covert* retrieval conditions all led to better performance on the final test than the *study once* condition, they were not significantly better than the *restudy* condition. We hypothesized that departures from many previous studies on testing — especially requiring JOLs in all conditions — may have been responsible for this unusual occurrence. However, first test performance was about 67% correct, so lack of feedback may also have been an issue, as subjects were not exposed to 33% of the items during the first test (but were, of course, exposed to 100% of the items in the restudy condition). Thus, Experiment 2 featured two procedural changes– subjects no longer made JOLs after each item presentation (except in the covert retrieval condition) and feedback was provided after each retrieval attempt. With these

modifications, subjects showed the standard testing effect; any condition where subjects attempted retrieval led to better performance on the final test than in the *restudy* and *study once* control conditions. Further, this experiment showed that responding to an item via typing or speaking led to better performance on the final test than the *covert* retrieval condition did, where subjects only made the delayed, cue-only JOL. However, subjects may not always covertly retrieve items while making a JOL, so in Experiment 3 we implemented the delay timing procedure to encourage subjects to retrieve the response covertly. In this experiment, final test performance was similar whether subjects responded overtly or covertly.

We can reach several conclusions from these results. First, response modality, in comparing typing to speaking, appears to have very little, if any, effect on retrieval, either during an initial test or in influencing future test performance. In fact, in both Experiments 1 and 2, the only significant difference between the *aloud* and *type* conditions was on the first test in Experiment 2 where subjects in the *type* condition did slightly better than those in the *aloud* condition. However, during the first test of Experiment 1, and on the final tests for both Experiments 1 and 2, there were no differences between speaking and typing, suggesting that both forms of responding have similar effects on retrieval. Subjects generally recalled the same amount regardless of whether they responded orally or by typing. Regarding the question of whether writing may lead to better performance on a second test (from subjects being allowed an additional study opportunity after typing their answers), the answer is no: writing and speaking led to equivalent testing effects.

If speaking and typing as forms of overt retrieval are equivalent, how do they influence retrieval on a future test compared to covert retrieval? Experiments 1 and 3 suggested covert retrieval is the same as overt retrieval, whereas Experiment 2 showed there was a difference. In Experiment 2, the covert retrieval condition led to worse performance on the final test than the overt retrieval conditions, whereas in Experiments 1 and 3 covert and overt retrieval conditions led to equivalent performance on the final test. As mentioned earlier, one of the challenges of investigating covert retrieval mechanisms is determining whether subjects are actually retrieving when making a delayed JOL. The timing procedure from Experiment 3, however, essentially forced subjects to covertly retrieve the item, and thus may be a better procedure for examining covert retrieval.

We are still left with a puzzle: How then to explain the differences in recall between Experiments 1 and 2? Although comparing between experiments must always be done with caution, it appears that the *covert* condition yielded similar levels of recall in Experiments 1 and 2 (56% and 59% respectively), whereas the *type* and *aloud* conditions showed a difference between experiments (58% and 68% for *type* and 58% and 69% for *aloud*), probably because feedback was provided in all conditions in Experiment 2. One possibility is that in Experiment 1, because subjects made JOLs after all items, they learned to thoroughly retrieve items prior to making JOLs. In Experiment 2, subjects only made JOLs in the *covert* retrieval condition. JOLs are an unusual procedure for most subjects, so practicing retrieval before making JOLs in Experiment 1 may have better encouraged subjects to retrieve items before making the JOL in the *covert* condition. Thus, in Experiment 2 subjects may not have been covertly retrieving items

before making the JOL as they were encountering the procedure for the first time. This supports the position that JOLs are not a valid substitute for true covert retrieval. Indeed, other research in our lab (Smith, 2011) utilized a similar timing procedure to that used in Experiment 3 and reached a similar outcome: covert and overt retrieval led to similar performance on a final test. Thus, covert retrieval appears to be just as effective as overt retrieval *as long as subjects are truly attempting to covertly retrieve information.*

Theoretical Implications

When Tulving (1983) suggested that all forms of retrieval have similar effects on future retrieval, he appears to have been correct. Response modality does not appear to influence the size of the testing effect – just thinking about an answer leads to similar benefits to memory as an overt response. This outcome supports the notion that it is retrieval processes that are causing the testing effect, rather than some by-product of producing the answer. In their review of the testing effect literature Roediger and Karpicke (2006a) examined several theories of the testing effect including: (1) effortful retrieval, (2) encoding variability, and (3) transfer-appropriate-processing. All of these theories suggest in some way that the memory improvement seen after testing results from retrieval processes that occur during testing. Each of theses theories makes different predications about how response modality should influence retrieval, which are discussed below.

Effortful retrieval hypotheses of the testing effect posit that engaging in retrieval results in a deeper processing of the item, similar to the deep level of encoding used in the levels of processing experiments (Craik & Tulving, 1975). Bjork (1975) argued that engaging in a difficult retrieval should enhance future retrieval, just as deep semantic

encoding enhances retrieval. Pyc and Rawson (2009) among others, have reported evidence in support of this hypothesis (see also Carpenter & Delosh, 2006). Essentially, a difficult retrieval at a first test will lead to enhanced recalled on a later test compared to an easy retrieval. The effortful retrieval hypothesis does not make strong predictions about how response modality will affect retrieval. On the one hand, response modality may not impact retrieval at all, as production occurs after retrieval. Intuitively, a subject mentally retrieves an item, and then subsequently produces it. Because production occurs after retrieval, it should not influence the difficulty of retrieval, and thus should not affect the strength of the memory. Production is not a necessary consequence of retrieval, although it is often required in memory experiments. (One criticism of this argument is that the production effect experiments (MacLeod et al., 2010) do find memory benefits for production, yet reading a word aloud could also be considered a staged process, as subjects may read the word silently before producing it.) On the other hand, writing a response to a question is certainly harder than just thinking about the answer. The increased difficulty of responding should affect future memory performance for all of the same reasons that a more difficult retrieval would. Perhaps if the specific response modality was more difficult, such as having subjects write answers with their non-dominant hands, then future retrieval would be enhanced due to the increased difficulty of response production. In summary, effortful retrieval hypotheses probably do not make strong predictions about the effects of response modality on retrieval.

Another proposed theory of the testing effect is encoding variability, or the elaboration of retrieval routes (McDaniel & Masson, 1985; Martin, 1968). Encoding variability theories suggest that having multiple access routes to a memory will result in

enhanced recall. An example of variable encoding would be asking someone to remember a word by thinking about the look, sound, and meaning of word. Processing multiple attributes of an item, therefore, should lead to better recall because many properties of the item were encoded. Theories of encoding variability suggest that varying response modality should lead to a variety of retrieval routes: a word presented visually, then recalled orally, then finally tested in writing should have several different access routes and should be easy to recall. Our data, however, do not support this view; response modality on a first test had almost no effect on performance on a second test. This is particularly interesting because Gardiner et. al. (1977) suggested that subjects should have enhanced memory for items that they recalled both in writing and by speaking, as the different response modalities should help subjects encode multiple attributes of the item. An experiment currently underway in our lab is a replication of Experiment 2, with the addition of three intermediate tests before the final test. Subjects respond in the same response modality across all three tests, or the response modality is varied across the three tests, providing a closer look at how encoding variability in response modality may influence retrieval. This research may reveal some further support for the encoding variability hypothesis, but as the current experiments did not yield any conclusive support for the influence of response modality on retrieval, it seems unlikely.

Finally, transfer-appropriate-processing theories (TAP; Morris, Bransford, & Franks, 1977), suggest that retrieval is enhanced when the processes engaged in at encoding are similar to the processes engaged in at retrieval. This theory is intuitively applicable to the testing effect – after all, what is the best way to prepare for a test other

than to practice taking the test? TAP suggests that responding in the same modality at a first and final test would yield the best recall (one of our predictions in Experiment 1). This is at odds with the prediction suggested by the encoding variability principle above (where engaging with a variety of response modalities would enhance recall), but nonetheless, TAP predicts that speaking at a first test should lead to better performance if the final test is oral rather than written or typed. As before, however, we did not find any indication that consistent response modalities across tests led to enhanced recall.

When considering these three theories of the testing effect in light of the current results, the predictions of the transfer-appropriate-processing and encoding variability principles are not supported. With regard to the to effortful retrieval hypothesis, the data are not particularly conclusive one-way or another. If all three theories would seem to make predictions, how come the results are not supportive of those principles? One possibility for why our data do not appear to support any of these theories is that any variation or benefit that may arise from a match or mismatch of response modalities is overpowered by the benefits of retrieval. Although some research (MacLeod et al., 2010 for example) suggests response modality can influence recall, retrieval practice effects may just overshadow any differences that arise from response modality. Another related possibility is that any enhanced distinctiveness that results from production in some form or another is lost before the final test two days later, whereas the benefits of retrieval practice will appear even larger on a delayed final test. One final possibility is that response modality does not affect retrieval processes at all, if only because the production of a retrieved item must necessarily occur after retrieval. As discussed above, if the retrieval process is entirely completed before production of the answer occurs, it seems

unlikely that the specific response modality could have an impact on the retrieval process per se.

There are several future directions for continuing exploration of response modality and its possible relation to retrieval. One simple approach, mentioned above, would be to have subjects practice retrieval multiple times before the final test. Taking three or more intermediate tests would likely result in a stronger manipulation, which may reveal differences between the different response modalities that may not appear after only one test. Multiple tests also would provide a vehicle for testing encoding variability ideas. Another approach would be to explore why production as an encoding manipulation appears to help on recognition tests, while production does not help when it occurs on a first test. Finally, examining various forms of retrieval by using more educationally relevant materials, such as geographical facts or more complex prose passages would add to the external validity of this research track.

Educational Applications

Several implications of this research are relevant to classroom education. For example, teachers often pose questions to their classes and call on a student to answer. If covert retrieval is just as effective as overt retrieval, than all of the students who attempt to mentally answer the question may be more likely to successfully answer a similar question on a future test – not just the student who is called on. A teacher could utilize a classroom procedure where she informs the students that she is about to ask a question and will call on one of the students at random, suggesting that every student should prepare to answer the question. In this way, asking questions to the class begins to

resemble the procedure from Experiment 3, thus building in more opportunities for retrieval practice throughout the year.

Another relevant application is in learning through flashcards. Some students believe that writing out flash cards by hand is the most important part of using flashcards. Although any additional study opportunity (and writing out all of the flash cards would certainly qualify) is beneficial, the benefits of retrieval practice do not appear to be tied to writing or speaking, suggesting that using pre-made flash cards may be just as effective as handwritten ones. Further, a student can just think about the answer before turning over the card, rather than saying the answer out loud. Indeed, future experiments could mimic a flashcard learning situation to more specifically address whether any form of overt retrieval will enhance future recall.

Conclusion

In summary the current experiments failed to find any evidence that response modality is a relevant factor in determining recall or in modulating the testing effect.. Speaking, typing, and covertly retrieving items all seem to lead to similar performance on a final test, suggesting that it is the act of retrieval that is critical in driving the testing effect rather than overt response production. Although this may contradict what some teachers and students intuit about memory, these individuals may take some solace in the finding that mental rehearsal appears to be an effective way to practice retrieval.

References

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604-616.

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., Pasher, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438-448.

Conway, M. A., & Gathercole, S. E. (1987). Modality and Long-Term Memory. *Journal of Memory and Language, 26*, 341-361.

Craik, F.I.M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Dunlosky, J. & Nelson, T. O. (1994). Does the sensitivity of Judgments of Learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545-565.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16,* 45-54.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6,* 1-104.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11*, 534-537.

Izawa, C. (1966). Reinforcement-test sequences in paired-associates learning. *Psychological Reports, 18,* 879-919.

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

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

Kimball, D. R. & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31,* 918-929.

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.

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review, 75,* 421– 441.

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "Delayed-JOL Effect." *Psychological Science, 2,* 267-270.

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

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

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

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

Schneider, W., Eshman, A., Zuccolotto, A. (2002). E-prime reference guide. Pittsburgh: Psychology Software Tools, Inc.

Smith, M. A. (2011). *Covert retrieval practice benefits retention as much as overt retrieval practice* (Unpublished master's thesis). Washington University in Saint Louis, Saint Louis.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315-316.

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

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 135-144.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465-478.