Mem Cogn (2013) 41:36–48
DOI 10.3758/s13421-012-0245-x

Does response mode affect amount recalled or the magnitude of the testing effect?

Adam L. Putnam · Henry L. Roediger III

Published online: 17 August 2012

Abstract The testing effect is the finding that retrieval practice can enhance recall on future tests. One unanswered question is whether first-test response mode (writing or speaking the answer) affects final-test performance (and whether final-test response mode itself matters). An additional unsettled issue is whether written and oral recall lead to differences in the amount recalled. In three experiments, we examined these issues: whether subjects can recall more via typing or speaking; whether typing or speaking answers on a first test can lead to better final-test performance (and whether an interaction occurs with final-test response mode) and whether any form of overt response leads to better final-test performance as compared to covert retrieval (thinking of the answer but not producing it). Subjects studied paired associates; took a first test by typing, speaking, or thinking about responses; and then took a second test in which the answers were either spoken or typed. The results revealed few differences between typing and speaking during recall, and no difference in the size of the testing effect on the second test. Furthermore, an initial covert retrieval yielded roughly the same benefit to future test performance as did overt retrieval. Thus, the testing effect was quite robust across these manipulations. The practical implication for learning is that covert retrieval provides as much benefit to later retention as does overt retrieval and that both can be effective study strategies.

Keywords Testing effect · Response mode · Retrieval practice

Thousands of memory experiments have been conducted using recall as the criterion variable, whether in cued recall, serial recall, or free recall. One factor that varies throughout the literature is response mode—whether subjects write or speak their responses. In the first studies of memory, only spoken responses were used (Ebbinghaus, 1885/1964), but we suspect that the majority of recall studies today use typed or written recall, for ease of scoring.

This article addresses two issues of response mode: how it may affect the amount recalled, and how it may affect later retention. Regarding the first—does response mode (speaking or writing) affect the amount recalled?—we can find few researchers who have seriously considered the question. Many studies have contrasted visual and auditory input, but few have investigated written versus spoken response mode. We suspect that most researchers assume that response mode does not matter, and thus use whatever procedure is handy for their purposes. As the bulk of the contemporary literature on long-term retrieval has predominantly used written or typed responding, it is important to address whether response mode can influence recall.

The second, and perhaps more interesting, issue is whether response mode on a first test (speaking or typing) affects the benefit of that test on a second test. That is, if subjects study material and are then asked to speak their responses or to type them on a first test, will this difference in response mode affect performance on a second test? And will the response mode of the second test matter in determining the benefits of the first test? These are the central questions driving our experiments.

Retrieval practice effects

Retrieval on a first test affects performance on a second test; this finding, called the *testing effect*, was first shown many years ago (Abbott, 1909; Gates, 1917), even if the attention

A. L. Putnam (✉) · H. L. Roediger III
Washington University in St. Louis,
St. Louis, MO, USA
e-mail: adam.putnam@wustl.edu

to the effect was not sustained (see Roediger & Karpicke, 2006a). In many situations, retrieval practice has been shown to be a more effective learning tool than restudying for an equivalent amount of time (Carrier & Pashler, 1992; Roediger & Karpicke, 2006b). A plethora of studies have shown that the testing effect occurs with different types of memory tests, over a wide range of materials, across different schedules of studying and testing, and with or without feedback (see Roediger, Putnam & Smith, 2011, for a recent review of the benefits of testing).

In the vast majority of testing effect experiments, subjects have written or typed their responses on both the initial and final tests. Does response mode (in particular, writing on both the first test and the second test) represent a necessary condition for the effect to occur? Perhaps writing or typing (rather than speaking) at a first test could boost performance on a second test because subjects are able to see their responses after they have written them, effectively increasing exposure time. Of course, this concern is stronger when subjects respond on paper and can scan their responses multiple times during the recall period. At any rate, according to this logic, writing or typing responses at a first test should yield a stronger testing effect than reporting answers verbally, due to extra study time. If this procedure is responsible (in whole or in part) for the benefits of retrieval practice, it could have important implications for our understanding of the testing effect. One present aim, then, is to determine whether response mode influences the testing effect. (As reviewed below, one can also use other theories to make a different prediction about the effects of response mode on memory—viz., that spoken retrieval would provide a greater benefit than written retrieval on a later test).

Response mode and memory

The relationship between modality and memory has been explored in the past, although research has typically focused on presentation modality (i.e., the modality effect; Harvey & Beaman, 2007; Penney, 1975) rather than on output or response modality. A few exceptions, however, do exist. Kellogg (2007) reported evidence that spoken recall yielded more idea units (but also more distortions) than written recall when subjects were asked to report what they remembered after reading or listening to the "War of the Ghosts" (Bartlett, 1932). Kellogg argued that because writing required the activation of additional graphemic codes, took longer than speaking, and was less practiced, it consumed more working memory capacity than did speaking, meaning that subjects had more resources available for recall when reporting answers verbally. Gardiner, Passmore, Herriot and Klee (1977) employed a two-test paradigm in which subjects wrote some words and spoke others during an initial test, but Gardiner et al. were interested in whether the subjects could recognize words that they had recalled earlier, rather than in the effects of response mode per se 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 did the groups that only wrote or only spoke their answers; there was no difference between the oral and written response groups. Gardiner et al. suggested that retrieval not just increases the strength of a trace, but also causes qualitative changes in the encoding (or recoding) of the trace. Retrieving an item and producing it orally, for example, results in both the 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. In sum, Gardiner et al.'s view suggests that different response modes affect recoding of memories in qualitatively different ways, which will subsequently affect future recall. If this is so, then first-test response mode should influence performance on a second test.

Other evidence, such as work on the production effect (Conway & Gathercole, 1987; Hopkins & Edwards, 1972; MacLeod, Gopie, Hourihan, Neary & Ozubko, 2010), has shown that saying a word aloud can enhance retention as compared to reading a word silently. MacLeod et al. suggested that producing (speaking) words creates a distinct verbal record that subjects can use to facilitate future recognition. This finding suggests that retrieval practice, at least as it is usually implemented, is always confounded with production. If production creates a more distinctive verbal record, then perhaps the testing effect may be due in part to having subjects overtly report their answers. Thus, overt retrieval (saying or writing the response) should produce a greater benefit than covert retrieval (thinking the response but not overtly producing it). Although the production effect has only been explored as an encoding manipulation, it is not unreasonable to assume that similar processes may occur during retrieval and may boost the testing effect. After all, retrieval processes may involve some form of encoding (McDaniel & Masson, 1985) or reconsolidation (e.g., Finn & Roediger, 2011). As a further point, much recent work in the embodied cognition tradition has suggested that involving action-oriented response modes may affect performance relative to pure thought (e.g., Wilson, 2002). If so, the motor output and kinesthetic feedback from speaking and writing might augment the testing effect relative to covert retrieval, which we will consider next.

Covert retrieval

Comparing covert retrieval to overt retrieval is perhaps the closest approximation to examining the effects of production

at retrieval rather than encoding, and the literature has mixed findings on whether there are differences between the two types of retrieval. Whitten and Bjork (1977) conducted an experiment comparing restudy, overt retrieval, and covert rehearsal. Their results led them to conclude that no spacing or testing effect occurred for the covert-rehearsal condition, in which subjects were instructed to mentally recall the word but not to produce it. Whitten and Bjork offered two explanations for the lack of improvement in the covert-rehearsal condition: (1) ". . . overt retrieval and covert rehearsal involved qualitatively different processes" (p. 472), or (2) a lack of experimental control over what subjects were doing during covert retrievals (i.e., perhaps they were not actually bringing the target items to mind). Regardless, Whitten and Bjork concluded that covert rehearsal did not have the same effect on memory as overt retrieval.

Carpenter, Pashler, Wixted and Vul (2008), however, showed that covert retrievals can enhance future test performance. Subjects who mentally retrieved the answers to obscure fact questions showed higher final-test performance than did a restudy control group. That is, they showed a testing effect with covert retrieval practice as compared to restudy. Even without comparison to an overt-response group, these results suggest that the testing effect can occur without an overt response during the first test. If so, the act of retrieval (bringing information to conscious awareness) primarily drives the testing effect, rather than response production (although overt responding may enhance the effect, as we discussed previously).

Izawa (1976) directly compared the effects of overt and covert retrieval on future memory performance in an experiment using a complex paired-associate learning design. Subjects experienced several cycles of studying and testing in which various conditions had different patterns of silent (covert retrieval) and vocalized (overt retrieval) trials intermixed with restudy trials. Although Izawa observed some subtle differences between conditions on the earlier trials, on the final test all conditions produced equivalent recall. She concluded that both vocalized and silent test trials had similar effects on future test performance.

A critical issue in comparing overt to covert retrieval is how the covert-retrieval trial is implemented. Subjects may or may not attempt retrieval on every trial, and the experimenter has no way of assessing compliance if the procedure simply instructs subjects to covertly retrieve. One solution is to use a task that requires subjects to retrieve the item without overtly producing it, such as making a delayed judgment of learning (JOL). In this procedure, subjects learn material (say, paired associates), and after a delay are given the cue member of the pair and are asked to predict whether they will be able to recall the target item on a later test (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991). Subjects' predictions are good (much better than in the standard JOL procedure, in which the cue and target are presented together at study), and some have argued that this outcome occurs because subjects attempt retrieval in order to complete the JOL. That is, if a subject can covertly retrieve an item, he will predict its successful recall in the future, whereas if he *fails* to covertly retrieve the item, he will predict not recalling the item in the future, because no feedback is given on the test (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). Because a delayed, cue-only JOL probably requires a covert retrieval, we used it in Experiments 1 and 2 in preference to more generic instructions telling subjects to "think of the answer." For Experiment 3, we used a different procedure that seems more effective in eliciting covert retrieval than delayed JOLs.

As reviewed above, the issue of how covert retrieval influences memory remains murky, with evidence suggesting both that covert retrieval should have equivalent effects on future recall and that overt retrieval will be more effective. We have also reviewed a pair of studies (Gardiner et al., 1977; Kellogg, 2007) that suggested that typing and speaking may influence memory in different ways. One alternative view is found in Tulving's (1983) general abstract processing system (GAPS) framework, which led him to argue that no difference should be observed between covert and overt retrieval. We will discuss this theory further in the General Discussion.

If response mode on the first test (spoken, typed, or covert) does influence the magnitude of the testing effect, such that overt production leads to a greater benefit on the final test than does covert production, then another prediction can be made. On the basis of ideas of transfer-appropriate processing (Morris, Bransford & Franks, 1977; Roediger, Gallo & Geraci, 2002), one might expect that first-test response mode might be differentially effective, depending on the second-test response mode. For example, if the final test requires vocal responding, vocal responding on the first test should confer greater benefits than would typed responding or covert retrieval. The contrary pattern might occur with typed responding (i.e., typed responding on the first test might lead to better performance on the final, criterial test if it also required typed rather than vocal responses). Of course, this predicted pattern only makes sense if response mode on a first test can be shown to have an effect on a second test, whatever the type of the second test.

Present research

In the present experiments, we examined whether first-test response mode (typed, spoken, or covert) affects the magnitude of the testing effect, as well as examining any differences in overall recall between typing and speaking on a

final test. For all experiments, we used a paired-associates procedure with a study phase, a first test, and then a second test. Response mode was varied at the first or final tests (or both). In Experiment 1, we compared how different response modes, such as typing, speaking, and making a JOL, influence future test performance. Experiment 2 was a replication with two procedural changes that yielded a much stronger testing effect. For Experiment 3, we introduced a new procedure with timing deadlines and different response options, to allow for a more direct comparison between the effects of covert and overt retrieval on later retention.

Experiment 1

Experiment 1 had three phases: an initial study phase, an intermediate phase in which response mode was manipulated on a first test, and a final phase in which response mode was manipulated on a second test. During the intermediate phase, subjects recalled words by speaking, by typing, or via a covert retrieval (making a JOL when they saw the cue word). In two control conditions, subjects either restudied the information (the *restudy* condition) or had no further exposure to it during the intermediate phase (the *study-once* condition). After each retrieval or restudy trial, subjects made a JOL for the current item, predicting their performance on the final test. In the covert condition, subjects made the JOL without producing the item, whereas in the two overt conditions, the JOL was made after the retrieval attempt. No feedback was given on the first test. Two days later, at the final test, subjects responded by either typing or speaking.

We had three predictions about final-test performance. First, as Tulving (1983) hypothesized in his GAPS framework, all forms of retrieval may have similar effects on memory; thus, his theory predicts that the aloud, type, and covert-retrieval conditions should all yield similar final-test performance. As those conditions are all different forms of retrieval practice, they should outperform the control conditions. A second hypothesis, derived from research on the production effect, suggests that making an overt response should enhance retention. Thus, the overt-response conditions, aloud and type, should boost recall more than the covert condition. If this outcome occurred, a third prediction could be made from the perspective of the transfer-appropriate processing framework (e.g., Morris et al., 1977; Roediger et al., 2002). If overt production produces a greater effect than covert retrieval, this framework would predict that matching response modes between the first and final tests (e.g., typing responses on both the first and final tests) would enhance performance relative to the conditions in which response mode was mismatched (e.g., typing responses on the first test and saying them aloud on the final test). Thus, besides answering a fundamental empirical question about the testing effect—whether response mode matters—these experiments may help delineate the most promising theoretical approaches to the testing effect.

Method

Subjects A group of 50 subjects from Washington University in St. Louis's research pool participated for course credit or payment ($10). Five of the subjects had incomplete data, either due to a computer error or because they failed to return to the lab for the second session, and they were replaced with five new subjects.

Stimuli Seventy-five weakly related word pairs were constructed (e.g., "airplane–trip" and "blossom–cherry"; Nelson, McEvoy & Schreiber, 1998). The pairs had a forward cue-to-target strength and a backward target-to-cue strength between .01 and .02 and were from three to nine letters long. The word pairs were within a medium range of concreteness and frequency and were all nouns.

Design and counterbalancing The experiment had a 5 (intermediate response mode: type, aloud, covert, restudy, or none—i.e., study once) × 2 (final-test response mode: aloud vs. type) mixed design. Intermediate response mode was manipulated within subjects, while final-test mode was manipulated between subjects (25 subjects in each group). The 75 word pairs were randomly divided into five lists of 15 pairs 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 the presentation order was counterbalanced across subjects. The final test manipulated response mode between subjects, with one group responding by speaking and one group responding by typing.

Procedure The experimental procedure consisted of a study phase, an intermediate test phase in which subjects were tested on some words and restudied others, and a final test phase. Subjects were tested individually. They were told that they would be learning word pairs and would take one test that day and another in two days.

In the study phase, subjects studied the 75 word pairs. The word pairs were presented in black, lowercase letters on a white background for 4 s, followed by a 500-ms interstimulus interval. All 75 word pairs were presented in random order, and the computer cycled through the entire list twice.

Afterward, subjects began the intermediate phase, in which they restudied some items and were tested on others. Items in four of the intermediate conditions were presented in blocks, whereas 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 6 s to type the target word into a box on screen. After 6 s, 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 0 (*no chance of recall*) to 100 (*absolutely sure I will recall it*). A blank screen appeared for 500 ms before the next trial began. 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 subjects were told to just read the cue word. After 6 s, the subjects made their JOL and moved on to the next trial. As we mentioned in the introduction, subjects likely covertly retrieve the target item in order to complete the delayed JOL (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). In the restudy condition, subjects saw both the cue and the target word during the 6-s window before making a JOL. The 15 words in the study once condition were not presented. After completing the intermediate phase, the subjects were dismissed with instructions to return to the lab in two days.

When subjects returned, they took a final cued recall test on all of the word pairs. A cue word was presented, and subjects were asked to generate the target word; there was no time limit. A total of 25 subjects responded by typing, and 25 responded by speaking. The typed responses remained on the screen until the subject pressed Enter, which began the next trial. Spoken responses were recorded by an experimenter, who then pressed Enter to begin the next trial. After finishing the cued recall test, subjects were thanked and debriefed.

Results

For all experiments obvious misspellings of a target word—"oyester" instead of "oyster"—were coded as correct. The JOL data are omitted because they are not germane to the main issues.

First-test performance Measures of recall were only available for the type and aloud conditions on the first test, as the other three conditions did not yield overt responses. Subjects recalled .68 (*SEM* = .04) in the type condition and .66 (*SEM* = .04) in the aloud condition, indicating that response mode did not influence performance on the first test. A *t* test confirmed that this difference was nonsignificant $t(49) = 0.926$, $p = .359$. Assuming a moderate effect size ($d = 0.50$), we obtained a power of .93. Subjects failed to recall over 30 % of the items in the overt-testing conditions, which has implications for interpreting the final-test results.

Final-test performance Table 1 shows the proportions of words correctly recalled on the final cued recall test. Remarkably, all conditions produced similar levels of recall, aside from the study-once condition. A 5 × 2 repeated measures ANOVA with intermediate-phase response mode as a within-subjects variable and final-test response mode as a between-subjects variable revealed a significant main effect of response mode at the first test, $F(4, 192) = 25.98$, $p < .001$, $\eta_p^2 = .35$. All four reexposure conditions (restudy or test) during the intermediate phase led to better performance on the final test relative to the study-once condition. There were no further differences, however, as the type, aloud, covert, and restudy conditions all led to similar performance on the final test. There appeared to be an effect of final-test response mode ($M = .56$ for the aloud group and $M = .45$ for the typed group, collapsed across intermediate-phase modes), but the outcome was not statistically significant, $F(1, 48) = 3.84$, $p = .056$, $\eta_p^2 = .074$; more importantly, the effect was not replicated in later experiments. There was no significant interaction between the initial test condition and the response mode on the final test $F(4, 192) = 0.25$, $p = .913$, $\eta_p^2 = .005$. Paired *t* tests revealed that words from the study-once condition were recalled significantly less well during the final test than were the words in any other condition. In other words, the type, aloud, covert, and restudy conditions all yielded higher performance than the study-once condition but did not differ significantly from one another. A post hoc power analysis assuming a moderate effect size ($d = 0.50$) and alpha equal to .05 yielded an achieved power of .93 for each comparison. The *t* test values comparing the reexposure conditions to the study-once condition were as follows: type, $t(49) = 12.42$, $p < .001$, $d = 1.15$; aloud, $t(49) = 11.83$, $p < .001$, $d = 1.10$; covert, $t(49) = 10.79$, $p < .001$, $d = 1.06$; and restudy $t(49) = 5.60$, $p < .001$, $d = 1.07$.

Table 1 Experiment 1: Mean recall on the final test after experiencing different response modes during the first test, relative to the response mode at final test

Final-Test Mode	Type	Aloud	Covert	Restudy	Study Once
Type	.51 (.05)	.52 (.06)	.49 (.05)	.51 (.05)	.24 (.05)
Aloud	.65 (.06)	.63 (.06)	.62 (.05)	.60 (.05)	.33 (.05)
Both groups	.58 (.05)	.58 (.06)	.56 (.06)	.55 (.05)	.29 (.05)

Standard errors of the means are in parentheses.

Discussion

Overall, the results of Experiment 1 suggested that response mode had little to no effect on retrieval. First, there were no significant differences between the number of items that subjects recalled either via typing or speaking on the first or final test. Second, the response mode 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 the study-once condition). Finally, covert retrievals appeared to be just as effective as overt responding in generating a testing effect. Our results showed no differences between responding orally, by typing, or even covertly retrieving, in terms of their effect on final-test performance.

One troublesome finding, however, is that at the final test, the restudy condition showed performance equal to the retrieval conditions (aloud, type, and covert). Previous research (e.g., Roediger & Karpicke, 2006b) had suggested that retrieval practice should benefit performance on the final test more than restudying, which is how some researchers currently define the testing effect. Of course, the original definition of the testing effect compared a testing condition against a study-once condition (e.g., Wheeler & Roediger, 1992), so by that criterion we did find a testing effect.

One possible reason for the weak testing effect is that no feedback was provided after retrieval attempts. Thus, in the testing conditions, subjects were only reexposed to the word pairs they could correctly recall (67 % on average), whereas in the restudy condition, subjects restudied 100 % of the pairs. This difference in exposure to pairs favoring the restudy condition could partly account for the weakened testing effect (see Wenger, Thompson & Bartling, 1980, and Kang, McDermott & Roediger, 2007, for analyses of this point).

A second issue that may account for the lack of a testing effect is the unexpectedly good recall in the restudy condition; the restudy condition produced nearly 100 % improvement in recall relative to the single-study condition (.55 vs. .29), which is unprecedented in the literature. In most experiments, a single restudy session yields more modest dividends. What differed in our experiment from the typical case? One possibility is that subjects were required to complete JOLs after restudying in our experiment, rather than simply to restudy under the same condition as in the initial trial. Making a delayed JOL with only the cue presented enhances future recall, because subjects likely engage in retrieval; but making a JOL with the cue *and* the target displayed may not enhance future recall, because the subject does not need to retrieve the target (Kimball & Metcalfe, 2003). However, it is still possible that subjects engaged in some sort of retrieval during the JOL (e.g., looking at the cue word and trying to recall the response before reading it). If performing JOLs after restudying caused more elaborative processing than usually occurs during restudying, recall may have been greatly boosted on the final test, and thus the testing effect (relative to the restudy-with-JOL baseline) was eliminated. Experiment 2 addressed both the JOL and feedback issues.

Experiment 2

In Experiment 2, we introduced two procedural changes in hopes of obtaining a stronger testing effect. The first change eliminated the JOL procedure after cue presentation in the type, aloud, and restudy conditions. Although previous research (Kimball & Metcalfe, 2003; Nelson & Dunlosky, 1991) has suggested 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, the results of Experiment 1 suggested that the restudied items may have been enhanced by subjects making a JOL, because performance increased so much beyond one presentation. Experiment 2 still included a covert condition in which subjects were provided with a cue and subsequently made a JOL, to investigate again whether covert retrieval produced performance equivalent to that following overt retrieval (typed or spoken aloud); however, JOLs were eliminated from the other conditions.

The second change was providing feedback during the type, aloud, and covert conditions. Providing feedback allowed subjects to correct any mistakes from the first test and reexposed all of the word pairs. Feedback has been shown to be effective in enhancing benefits resulting from retrieval practice on a delayed final test (Butler & Roediger, 2008). Thus, the type, aloud, and covert conditions should show an increase in final-test performance relative to the restudy condition. Of course, if we found a testing effect relative to the restudy condition in Experiment 2, we could not know which factor (no JOLs or feedback) led to the difference. However, our aim was to find the testing effect under standard conditions (which often involve feedback and, in testing experiments, do not involve making JOLs).

Thus in Experiment 2 subjects studied paired associates in a first phase, and during an intermediate phase they recalled some words by typing or speaking, made a JOL for some words, and restudied others (with other words not being exposed during this phase). Two days later, the subjects returned to the lab and took a final cued recall test that varied between groups (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 remained the same.

Method

Subjects and materials A group of 50 subjects from the same pool as in Experiment 1 participated for course credit or cash. Six of the subjects failed to return to the lab for the second day of testing and were replaced with six new subjects. The same 75 weakly related word pairs were used as in Experiment 1.

Design and counterbalancing As before, the experiment had a 5 (intermediate response mode: type, aloud, covert, restudy, or study once) × 2 (final-test response mode: aloud vs. type) mixed design. First-test response mode was manipulated within subjects, while final-test mode was manipulated between subjects. The 75 word pairs were randomly divided into five lists of 15. The five lists of word pairs were rotated through the five conditions in the intermediate phase (type, aloud, covert, restudy, and not presented). The response mode at final test was varied between subjects, being either entirely typed or entirely oral, with 25 subjects in each condition.

Procedure The experimental procedure closely mirrored Experiment 1. During the study phase, subjects were presented with the weakly related word pairs and saw the entire list twice. In the intermediate phase, subjects were tested on some words and restudied others. As before, in the type and aloud blocks, subjects were presented with the cue word and had 5 s to respond by appropriately typing or saying the target word. The correct answer was then displayed for 2 s, and subjects moved on to the next trial. Subjects did not make a JOL after attempting retrieval of the target word and receiving the correct-answer feedback. However, in the covert-retrieval block, subjects saw the cue word and made a JOL. They had 5 s to respond, and then the correct response was displayed for 2 s after making the JOL. In the restudy condition, subjects saw both the cue and the target word for 7 s, equating the total exposure time per item with the retrieval conditions (without making a JOL). Finally, one set of 15 words was not presented during the intermediate phase: the study-once control condition. Subjects then left the lab and 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. As in Experiment 1, all subjects were tested individually with an experimenter present at all times.

Results

First-test performance Unexpectedly, subjects performed better in the type condition (M = .70, SEM = .04) than in the aloud condition (M = .60, SEM = .04). A paired samples t test, $t(49) = 2.78$, $p = .008$, $d = 0.33$, revealed that the difference was significant. Because subjects received feedback on all trials, however, exposure differences were minimized. The direction of the difference was opposite that of the trend in the final test in Experiment 1.

Final-test performance Table 2 shows performance on the final cued recall tests, broken down by response mode at the first and final test. The overt-response conditions produced a larger testing effect than did the covert-retrieval (JOL) condition, but all three conditions were elevated relative to the study-once or repeated-study baseline conditions, thus revealing a testing effect. A 5 × 2 repeated measures ANOVA with intermediate-phase response mode as a within-subjects variable and final-test response mode as a between-subjects variable revealed a main effect of intermediate-phase response mode, $F(4, 192) = 80.05$, $p < .001$, $\eta_p^2 = .63$. There was no main effect of response mode at the final test, $F(1, 48) = 1.28$, $p = .26$, nor was the interaction significant, $F(4, 192) = 1.41$, $p = .23$, Since there was no effect of final-test response mode on performance, those two groups were combined for the remaining analyses.

Planned comparisons revealed several effects of intermediate-phase response mode on final-test performance. First, we found no significant difference between the type and aloud conditions, although both of those conditions yielded significantly higher recall than the covert condition [aloud vs. type, $t(49) = -0.40$, $p = .69$; type vs. covert, $t(49) = 3.91$, $p < .001$, $d = 0.34$; aloud vs. covert, $t(49) = 4.39$, $p = .001$, $d = 0.39$], the restudy condition [type vs. restudy, $t(49) = 7.99$, $p < .001$, $d = 0.96$; aloud vs. restudy, $t(49) = 8.54$, $p < .001$, $d = 1.04$], and the study-once condition [type vs. study once, $t(49) = 12.19$, $p < .001$, $d = 1.45$; aloud vs. study once, $t(49) = 15.38$, $p < .001$, $d = 1.55$]. Second, the covert condition yielded significantly higher recall than did the restudy and study-once conditions [covert vs. restudy, $t(49) = 5.52$, $p < .001$, $d = 0.64$; covert vs. study once, $t(49) = 11.88$, $p < .001$, $d = 1.13$]. Finally, the restudy condition yielded higher recall than did the study-once condition [restudy vs. study once, $t(49) = 4.97$,

Table 2 Experiment 2: Mean recall on the final test after experiencing different response modes on the first test, relative to the response mode at final test.

Final-Test Mode	Type	Aloud	Covert	Restudy	Study Once
Type	.69 (.06)	.70 (.05)	.62 (.05)	.49 (.05)	.37 (.06)
Aloud	.66 (.05)	.67 (.04)	.56 (.05)	.37 (.04)	.26 (.03)
Both groups	.68 (.04)	.69 (.03)	.59 (.03)	.43 (.03)	.31 (.03)

Standard errors of the means are in parentheses.

$p < .001$, $d = 0.48$]. Note that without the JOL procedure, the restudy condition showed only a modest benefit as compared to the study-once condition, unlike in Experiment 1. Again, a post hoc power analysis was conducted—assuming a moderate effect size of 0.50 and a sample size of 50—that revealed achieved power of .93. To summarize, the aloud and type conditions yielded the best recall on the final test, followed in order by the covert condition, the restudy condition, and the study-once condition.

Discussion

The results of Experiment 2 showed a stronger testing effect as compared to Experiment 1: The three retrieval conditions led to higher performance than in the restudy and study-once conditions. The overt-response conditions (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 final-test retrieval was not as strong as with overt retrieval. This outcome suggests that testing effects may be driven both by the act of retrieval and by the production of an answer, yet this pattern of results differed from that in Experiment 1, in which covert retrieval produced effects equivalent to those of overt retrieval. One possible reason for this difference between the experiments is that providing feedback in Experiment 2 may have undermined subjects' motivation to covertly retrieve the items (knowing that they would see the correct answer in a few seconds may have led to their not trying to retrieve the answer on all test trials). However, the covert condition did lead to better performance than the restudy condition, implying that on some trials subjects *were* engaging in covert retrieval.

Providing feedback in Experiment 2 generally increased performance on the final test relative to Experiment 1. 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 have affected all of the retrieval conditions, the difference in outcomes between experiments was probably due to the absence of JOLs in Experiment 2. As mentioned earlier, subjects showed a large improvement from the study-once condition to the restudy condition in Experiment 1 (.29 to .55), but had a much smaller (and more typical) improvement in Experiment 2 (.31 vs. .43). Apparently, the additional JOL during the restudy phase in Experiment 1 may have engaged additional processing that benefited retention.

The finding that the overt-response conditions yielded better recall than the covert-retrieval condition is consistent with some earlier research (e.g., Izawa, 1976), but it differs from the results in Experiment 1 and from other, earlier research (Carpenter et al., 2008). Why should the covert condition not be as effective as the overt condition? One possibility, noted above, is that the JOL procedure may not be a perfect substitute for covert retrieval, especially with immediate feedback, as in Experiment 2. Although most researchers have argued that a delayed JOL involves covert retrieval of a response, subjects may sometimes make such judgments on the basis of other factors, such as the familiarity or fluency of the cue word (Dunlosky & Metcalfe, 2009, pp. 104–110). If so, subjects may not always attempt covert retrieval, likely due to the expectation of immediate feedback, which may result in the inferior performance in the covert condition as compared to the overt conditions. To overcome these problems, in Experiment 3 we employed a new procedure to elicit covert retrievals (described below) by asking subjects to retrieve the item before knowing whether or not they would need to produce it.

The issue of whether speaking or typing responses during recall leads to the same number of total items recalled remains murky. Experiment 1 showed a trend on the final test for spoken recall to lead to greater performance than written recall. In Experiment 2, however, the reverse pattern occurred on the first test, and also trended in that direction on the final test (a 7 %, albeit not significant, difference). We suspect that these variations are simply noise and that the two procedures do not lead to overall differences in the number of items recalled. We will consider this issue at greater length in the General Discussion.

Experiment 3

Experiment 3 was designed to further investigate the relationship between covert and overt retrieval and their influences on later tests by using a procedure that permitted greater control over covert retrieval than does requiring a JOL. In our procedure, subjects saw a cue word, but could not respond; after 4 s, either the word "Recall!" appeared and the subjects were asked to report the target word, or the phrase "Do you remember the target word?" appeared, and subjects were asked to report whether or not they could remember the target. Subjects had only 1.5 s to respond. The timing procedure effectively forced subjects to covertly recall the word initially (if possible) without knowing whether or not they would need to report it. We judged that this procedure would permit a cleaner comparison between covert and overt retrieval than did the JOL procedure.

Method

Subjects and materials A group of 25 subjects from Washington University in St. Louis's subject pool participated for cash or course credit. Six of the subjects failed to return for

the second day of the experiment and were subsequently replaced. The same materials were used as in Experiment 1.

Design In this experiment, we 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 (the *restudy* condition). One list of items was not presented during this phase, the *study-once* control condition. Unlike in the previous experiments, the final test was typed for all subjects.

Procedure The experiment consisted of a study phase, an intermediate phase, and a final cued recall test two days later. Subjects first completed a short practice phase allowing them to rehearse the procedure and to ask questions. During the study phase, 64 word pairs were presented for 3 s each in a random order. Subjects played a video game for 2 min 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). Sixteen word pairs were not presented during this phase, those in the study-once control condition. Unlike in the previous experiments, where the different conditions were blocked together, the different trial types were mixed in Experiment 3 to prevent subjects from knowing until the last moment whether they would actually report the target item.

On the restudy trials, the cue and target words (e.g., "airplane–trip") were displayed for 7.5 s. In the retrieval conditions, however, the cue word appeared with a series of question marks ("airplane–?????"). Subjects were instructed to bring the target word to mind, if possible, during the 4-s window, but were not to make any indication whether or not they had succeeded in doing so. After 4 s, 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 s to respond. In both cases, an experimenter recorded response accuracy, 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 2 s. A 1-s interstimulus interval indicated the start of the next trial.

After finishing the intermediate phase, the 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 typed and had no time limit.

Results

First-test performance On the first test, subjects recalled .42 (*SEM* = .04) of the word pairs correctly in the overt-retrieval condition. Comparatively, in the covert condition, subjects self-reported remembering .51 (*SEM* = .04) of the word pairs. Because one condition was an overt retrieval, while the other was 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, which is not surprising, given that subjects are often overconfident in their metamemory judgments (Dunlosky & Nelson, 1994). A paired-samples *t* test, $t(24) = -2.28$, $p = .03$, $d = 0.41$, revealed that this difference was significant.

In the covert condition, subjects were reporting whether they thought that they could accurately recall the target item. These subjects were 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 remembered 74 % of the items that they claimed to remember at the first test and 25 % of the items that they claimed to not remember. The latter value may seem high, but feedback was provided after the initial test.

Final-test performance Figure 1 shows the proportions of items recalled on the final test. Clearly, overt and covert retrieval produced similar levels of performance, and both conditions were superior to the study-once and restudy conditions. A one-way repeated measures ANOVA revealed a significant effect of intermediate-phase processing, $F(3, 72) = 46.23$, $p < .001$, $\eta_p^2 = .66$. Planned pairwise comparisons showed that both the overt- and covert-retrieval

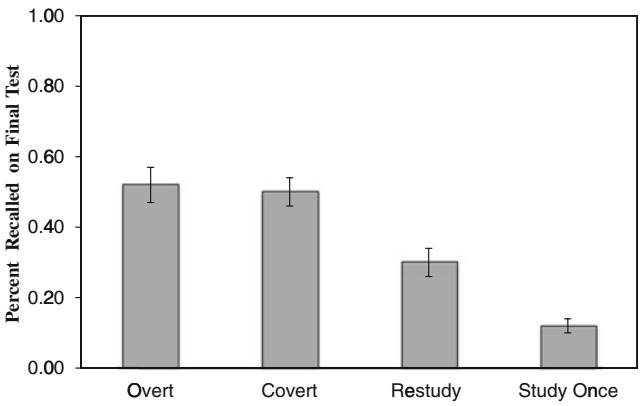

Fig. 1 Proportions of items recalled on the final test as a function of processing condition during the intermediate test phase in Experiment 3. Error bars represent the standard errors of the means.

conditions led to better performance on the final test than did the restudy condition, $t(24) = 5.72$, $p < .001$, $d = 0.92$, and $t(24) = 4.86$, $p < .001$, $d = 0.91$, and than did the study-once condition, $t(24) = 9.69$, $p < .001$, $d = 2.02$, and $t(24) = 8.72$, $p < .001$, $d = 2.15$, respectively. In other words, a testing effect was found, regardless of whether our baseline comparison was the study-once condition or the restudy condition. Furthermore, the restudy condition produced better recall on the final test than did the study-once condition, $t(24) = 5.33$, $p < .001$, $d = 1.07$. Finally, and most importantly, there was no significant difference between the overt- and covert-retrieval conditions, $t(24) = 0.41$, $p = .685$. A post hoc power analysis (assuming a moderate effect size: $d = 0.50$) revealed power equal to .67. Apparently, when the covert-retrieval condition is made quite similar to the overt-retrieval condition, no difference in the benefit of testing occurs in paired-associate learning.

Discussion

Experiment 3 was designed to determine whether making an overt retrieval had the same effect on future recall as retrieving an item under conditions designed to encourage covert retrieval. No difference emerged, although of course such a conclusion involves accepting the null hypothesis. This outcome differs from that in Experiment 2, in which JOLs were used to encourage covert retrieval. The testing procedure in Experiment 3 seemed more likely to force subjects to bring items to mind, because they were only required to produce the item half of the time. Subjects did report covertly retrieving the items (even at a greater rate than actually occurred in the overt-retrieval condition), so we expect that they followed the directions.

On the final test, subjects recalled words at similar levels in the covert and overt conditions: Both led to higher recall than did the restudy and study-once control conditions, indicating a testing effect 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. Of course, since feedback was provided on every trial, it is possible that a test potentiation effect occurred as well (Arnold & McDermott 2012; Izawa, 1966).

General discussion

The basic purpose of these experiments was to answer several interrelated questions about response mode at test: First, would response mode (typed or aloud) lead to differences in overall recall in paired-associate learning? The answer seems to be no, in that no consistent advantage emerged from one mode relative to the other. No difference occurred during the first test in Experiment 1, but on the final test, subjects who spoke aloud recalled about 10 % more words than those who typed (although this difference did not reach significance). On the first test in Experiment 2, the opposite pattern occurred, but on the final test there was no difference (with a nonsignificant advantage favoring typed responses). Collapsing across Experiments 1 and 2 and across the first and final tests (and weighting by the number of observations) yielded a mean of .56 for both the aloud and typed conditions, suggesting that there were no differences in the amounts recalled. Given that the aggregated means were the same and the inconsistencies across experiments, we conclude that no difference occurs in total recall from typing or speaking responses, at least in paired-associate learning.

A second question was whether greater testing effects would occur from spoken or written responding on a first test when measured on a second test. We also asked, third, whether the final-test response mode would interact with the first-test response mode. The answers to both of these questions were negative. The results from Experiments 1 and 2 consistently showed equivalent performance on the delayed test after typed or spoken recall on the first test. In addition, the same pattern occurred, no matter how recall was tested, such that matching of response modes between the two tests did not affect the number of items recalled. Of course, these conclusions again rest on failing to reject the null hypotheses, but there was no hint of an effect for either the main effect of spoken versus typed testing in the first test on recall in the second test, nor for the interaction of first-test mode and second-test mode. Furthermore, the results were consistent across two experiments. We concluded that equivalent testing effects occur from written and spoken recall in paired-associate learning.

A fourth issue was whether overt retrieval would produce a greater testing effect than did covert retrieval. Again, we concluded that the answer was no, although the results of Experiment 2 were inconsistent with this claim. The results of Experiment 3 are most telling in this regard, because the procedure there was most likely to have equated overt and covert retrieval on all features except response mode. The results from Experiment 2, which showed a weaker testing effect from covert retrieval than from overt retrieval, used a JOL procedure to try to encourage covert retrieval. Although JOLs sometimes require covert retrieval (e.g., Kimball & Metcalfe, 2003), our version of the procedure in Experiment 2 may have minimized covert retrieval due to providing feedback, as discussed earlier. Experiment 1, which used JOLs without feedback to induce covert retrieval, supported this notion: The covert-retrieval condition produced testing effects equivalent to those in the aloud and type conditions. Furthermore, another set of experiments by Smith and Roediger (2011)—ones that used rather different methods from those used here—led to the same

conclusion, that overt and covert retrieval produce testing effects of the same magnitude.

In addition to providing answers to four basic questions about response mode and its effects, we discovered one other interesting tidbit. When subjects have studied a word pair and then restudy it, the restudy phase has a greater impact when subjects make a JOL on the second study trial (Exp. 1) than when they do not (Exp. 2). The advantage of restudy over study once was .27 in Experiment 1, with the JOL, but only .12 in Experiment 2, without the JOL, $t(98) = 2.818$, $p < .001$ (although, of course, this comparison was across experiments). We believe that the use of the JOL procedure in Experiment 1 accounts for why we did not find a testing effect relative to the restudy baseline, as has often occurred in similar experiments (Carrier & Pashler, 1992).

Theoretical implications for the testing effect

We outlined several theories or phenomena that led us to expect that different response modes during an initial test may lead to differential performance on a later test. Likewise, we expected that overt production during an initial test would enhance performance on a later test relative to covert retrieval. First, MacLeod et al. (2010) showed that production during initial encoding could enhance memory relative to silent reading, so we predicted that producing an item during an initial test might also produce distinctive cues that led to greater retention than did covert retrieval. Bolstering this prediction is work from the embodied cognition literature (e.g., Wilson, 2002): Writing a word produces kinesthetic cues and other forms of muscular feedback that might provide a distinctive encoding; similarly, speaking a word aloud should provide distinctive articulatory and auditory cues relative to covert retrieval. However, despite these various predictions, typing and speaking the target word during the first test did not differentially affect later performance in Experiments 1 and 3. Neither response mode produced any benefit relative to covert retrieval.

We had also predicted a possible interaction between first- and final-test response modes. According to the transfer-appropriate processing framework (e.g., Roediger et al., 2002), performance on a criterial test should benefit more if that test requires or is consistent with the form of earlier encoding. Thus, for example, if speaking on the first test had led to greater responding on the second test than either covert retrieval or typing, one might have expected the effect to be greater if the second test were also accomplished by speaking rather than typing. However, because there was no main effect of first-test response mode, the possibility of such an interaction was slight, and obviously did not occur. Because the manipulation was not effective, the transfer-appropriate processing prediction was null and void.

If these theories led to predictions of larger testing effects for overt than for covert responding, why were these predictions not borne out? One possibility that we must acknowledge is that our conclusions are based on failures to reject the null hypothesis, which is never the strongest way to make inferences. In our defense, however, there was no hint of a difference between spoken and typed response modes in Experiments 1 and 2, and no effect of covert versus overt production in Experiments 1 and 3, and each experiment showed reasonable power in detecting a moderate effect size. (We argue that the somewhat discrepant results in Experiment 2 were probably artifactual: Subjects who received immediate feedback in that covert-retrieval condition might not have consistently tried to covertly retrieve). Thus, though they were null, we replicated both critical results in slightly different ways across experiments.

Another possibility is that the effect of retrieval practice was so strong (e.g., Karpicke & Roediger, 2008) that this variable overwhelmed any possible differences in response mode, especially from only a single test. If multiple tests were given during the first testing phase, perhaps effects of response mode might have been observed on the final test two days later.

Of course, one last possibility is that response mode simply does not affect the processes involved in the testing effect, if only because the production of a retrieved item must necessarily occur after retrieval. If the retrieval process is entirely completed before production of the answer occurs, as we tried to instantiate in Experiment 3, then the specific response mode simply has no impact on the changed nature of the memory trace after retrieval. This state of affairs may seem odd according to some accounts (reviewed above), but it is consistent with at least one theory: Tulving's (1983) GAPS framework. When specifying his theory, Tulving explicitly considered the mode of response and its resultant effects. He wrote 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" (Tulving, 1983, p. 47). Although he did not spell it out, Tulving's reasoning indicates that if retrieval processes are identical up to the point of production in "thinking about" the response, then making the response overtly will not alter the nature of the underlying representation and any changes already effected by retrieval. As we have seen from other theories, a counterargument can be made to this claim, but our results show that Tulving's hypothesis is correct.

Possible 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, then perhaps students can be trained to always bring an answer to mind in preparation of answering the question. If the retrieval attempt is successful, the direct benefit of testing will occur, and if the attempt is unsuccessful, the students may still indirectly benefit from test-potentiated learning when feedback is provided. If a teacher used the procedure of posing questions and then randomly calling on students to answer, perhaps all students would prepare to answer the question, and the gains would be realized. In this way, asking questions to the class begins to resemble the procedure from Experiment 3, requiring students to prepare a response, if they can. The typical alternative of asking a question and waiting to call on someone with a raised hand may not produce the same good effect if many students never bother to attempt retrieval.

Another relevant application is in learning through flashcards or similar devices. Although some students may believe that saying answers aloud before turning over a card is a key element of what makes flashcards work, our research suggests that just thinking about the answer is sufficient. This is a critical consideration if students are studying in a library or any other setting where they would not want to bother others by responding aloud. Furthermore, students today are beginning to use flashcard applications on mobile devices; covert retrieval avoids having to type responses on a small keyboard.

Conclusion

In summary, the present experiments failed to find any evidence that response mode is a relevant factor in determining either the total amount of recall or in modulating the magnitude of the testing effect. Speaking, typing, and covertly retrieving responses in paired-associate learning of word pairs all led to similar levels of recall 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 theories and many students intuit about how memory works, we may all take some solace in the finding that covert retrieval is an effective method of enhancing learning.

Author note This research constituted a masters thesis of the first author under the supervision of the second author.

We thank David A. Balota and Mitchell S. Sommers for serving on the committee and for their helpful comments. In addition, Jinkun Zhang, Benjamin Hoener, Paige Madara, Kate Margolis, and Kelly Young aided in the data collection. This research was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation and by a National Science Foundation Graduate Research Fellowship.

References

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

Arnold, K. M., & McDermott, K. B. (2012). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. doi:10.1037/a0029199

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36,* 438–448. doi:10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26,* 341–361. doi:10.1016/0749-596X(87)90118-5

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

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

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.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.)*. New York, NY: Dover. Original work published 1885.

Finn, B., & Roediger, H. L., III. (2011). Enhancing retention through reconsolidation: Negative emotional arousal following retrieval enhances later recall. *Psychological Science, 22,* 781–786. doi:10.1177/0956797611407932

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.

Harvey, A. J., & Beaman, C. P. (2007). Input and output modality effects in immediate serial recall. *Memory, 15,* 693–700.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11,* 534–537. doi:10.1016/S0022-5371(72)80036-7

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

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

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19,* 528–558.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Kellogg, R. T. (2007). Are written and spoken recall of text equivalent? *The American Journal of Psychology, 120,* 415–428.

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. doi:10.1037/a0018785

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. doi:10.1016/S0022-5371(77)80016-9

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

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.

Penney, C. G. (1975). Modality effects in short-term verbal memory. *Psychological Bulletin, 82,* 68–84. doi:10.1037/h0076166

Roediger, H. L., III, Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from levels of processing framework. *Memory, 10,* 319–332.

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 1–36). San Diego, CA: Elsevier/Academic Press.

Smith, M. A., & Roediger, H. L., III. (2011). *Covert retrieval practice benefits retention as much as overt retrieval practice*. Manuscript submitted for publication.

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, NY: 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.

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

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9,* 625–636. doi:10.3758/BF03196322