Journal of Experimental Psychology:
Learning, Memory, and Cognition
2013, Vol. 39, No. 6, 1712–1725

0278-7393/13/$12.00 DOI: 10.1037/a0033569

Covert Retrieval Practice Benefits Retention as Much as Overt Retrieval Practice

Megan A. Smith and Henry L. Roediger III
Washington University in St. Louis

Jeffrey D. Karpicke
Purdue University

Many experiments provide evidence that practicing retrieval benefits retention relative to conditions of no retrieval practice. Nearly all prior research has employed retrieval practice requiring overt responses, but a few experiments have shown that covert retrieval also produces retention advantages relative to control conditions. However, direct comparisons between overt and covert retrieval are scarce: Does covert retrieval—thinking of but not producing responses—on a first test produce the same benefit as overt retrieval on a criterial test given later? We report 4 experiments that address this issue by comparing retention on a second test following overt or covert retrieval on a first test. In Experiment 1 we used a procedure designed to ensure that subjects would retrieve on covert as well as overt test trials and found equivalent testing effects in the 2 cases. In Experiment 2 we replicated these effects using a procedure that more closely mirrored natural retrieval processes. In Experiment 3 we showed that overt and covert retrieval produced equivalent testing effects after a 2-day delay. Finally, in Experiment 4 we showed that covert retrieval benefits retention more than restudying. We conclude that covert retrieval practice is as effective as overt retrieval practice, a conclusion that contravenes hypotheses in the literature proposing that overt responding is better. This outcome has an important educational implication: Students can learn as much from covert self-testing as they would from overt responding.

Keywords: retrieval practice, testing effect, covert retrieval, memory

Research dating back a century has shown that taking a test is not a neutral assessment of memory (Abbott, 1909). Instead testing, or retrieval practice induced via testing, is a potent way to improve retention (see Roediger & Karpicke, 2006a for a review). Further research has also shown that retrieval practice can benefit retention in practical settings, such as middle-school classrooms (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011) and college courses (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). Because the direct effects of retrieval practice on retention are generally robust, cognitive psychologists have recommended that retrieval practice via testing be used as a way to promote learning in the classroom (McDaniel, Roediger, & McDermott, 2007; Roediger, Putnam, & Smith, 2011).

Researchers examining the retention benefits of retrieval practice have almost exclusively employed retrieval with overt responding. That is, during initial retrieval practice in these experiments, subjects are required to produce an overt response by writing, typing, or speaking. Covert retrieval—bringing information to mind or mentally rehearsing it—has rarely been used in prior research. This is of course not without good reason. Researchers are often interested in performance during retrieval because retrieval success is important for obtaining the positive effects of retrieval practice (see, for example, Butler, Marsh, Goode, & Roediger, 2006). If subjects do not produce an overt response during initial retrieval, then performance cannot be measured and the researcher cannot know how well subjects performed during initial retrieval practice. However, for both educational and theoretical reasons, the issue of whether covert retrieval provides as great an effect as overt responding is of interest. Consider the practical educational reason first. If students practice retrieval by self-testing as a study strategy then it clearly matters to them if overt responding produces greater retention than covert retrieval. Overt retrieval is more time consuming and requires a private space to study, and so if covert retrieval benefits retention just as much as overt retrieval does, students would be relieved of the necessity for overt responding.

There are theoretical reasons to think that overt retrieval may benefit retention more than covert retrieval. There may be a mnemonic benefit due to actually producing a response itself (rather than just holding it in mind) that contributes to the positive

This article was published Online First July 1, 2013.

Megan A. Smith and Henry L. Roediger III, Psychology Department, Washington University in St. Louis; Jeffrey D. Karpicke, Department of Psychological Sciences, Purdue University.

Megan A. Smith is now at Department of Psychological Sciences, Purdue University.

This research and the writing of this article were supported in part by a Collaborative Activity Grant from the James S. McDonnell Foundation. This article is based on Megan A. Smith's master's thesis submitted to Washington University in St. Louis under the direction of Henry L. Roediger III. We thank Mark McDaniel and Kathleen McDermott for serving on the thesis committee; Brittany Butler and Kelly Young for assistance with data collection and entry; James Nairne, Althea Bauernschmidt, Janell Blunt, Phil Grimaldi, and Melissa Lehman for comments on this article; and Greg Francis and Ludmila Nunes for assistance with the meta-analyses.

Correspondence concerning this article should be addressed to Megan A. Smith, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907. E-mail: smith598@purdue.edu

effects of practicing retrieval. For example, producing the items during retrieval practice may serve to make the items more distinctive and therefore more memorable later. Producing the words may add features to the item, such as the fact that a specific motor response was executed. At the time of recall, subjects will be required to construct a search set and discriminate among items in their search set that were a part of the original study episode and those that were not (Raaijmakers & Shiffrin, 1980, 1981). If the items that are overtly produced are made more distinctive by doing so, then it will be easier for the subject to discriminate those items among the others in the search set. Research on the production effect suggests that producing an overt response benefits memory for this reason. The production effect refers to the fact that producing a word out loud during study results in greater retention of that word relative to words that were only read silently during study, at least when within-subject, mixed list designs are used (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Although the overt response occurs during study in production effect experiments, the positive effect for spoken words over words read silently suggests that producing information overtly results in superior memory relative to covertly rehearsing information. In the production effect literature, overtly producing the words during study is thought to embellish or distinguish the produced words relative to those read silently. During retrieval practice, it is possible that overt responding might also make items more distinctive. For this reason, it is quite possible that overt retrieval practice will lead to better memory than simply practicing retrieval covertly.

Alternatively, it is possible that covert retrieval practice may benefit retention more than overt retrieval. Covert retrieval may be more difficult than overt retrieval because covert retrieval likely places additional demands on output monitoring especially in tasks such as free recall. During covert retrieval practice, one needs to monitor what has already been retrieved in their mind while they are also retrieving other relevant information. Some theories have suggested that processing difficulty can actually aid memory, when all other things are held equal (Bjork, 1999). The additional demands on output monitoring required by covert retrieval may produce a desirable difficulty. In addition, output monitoring during covert retrieval is unlikely to be perfect. Because monitoring will not be perfect, subjects may be more likely to mistakenly repeatedly recall items during covert retrieval conditions. If covert retrieval affords more repeated recalls, then learning should be improved simply because repeated retrieval practice greatly improves memory (Karpicke & Roediger, 2007, 2008).

Finally, it may be the case that both overt and covert retrieval produce equivalent benefits on learning. In other words, whether the information is produced overtly may not be a relevant dimension for retrieval practice effects. If the benefit of practicing retrieval arises because the subject is bringing a prior experience to mind (Karpicke & Zaromb, 2010), and both covert and overt retrieval require this process, then the two should produce equivalent benefits on learning and memory. If this is the case, then it would indicate that the mechanisms responsible for retrieval practice (e.g., establishing a search set, discriminating among other items within that set, or even elaborative processing during retrieval, Karpicke & Smith, 2012) do not depend on overt responding and do not seem to be influenced or altered by this dimension.

The experiments reported here were designed to examine overt and covert retrieval practice, and their effects on learning and memory. As was mentioned previously, most prior studies have used overt recall during retrieval practice. Some previous studies have shown that retrieval practice effects can also be obtained under covert retrieval conditions. For example, Carpenter and Pashler (2007) showed that a test involving covert retrieval improved visuospatial map learning. In their experiment, subjects studied maps containing a number of different features. During the initial test, subjects were given an incomplete version of the map and were instructed to form a mental image of any missing features. Covert retrieval was used here because producing overt responses during testing would not be possible or natural. Forming the mental image of the missing pieces resulted in a more accurate reproduction of the map later relative to restudying the map, indicating that covert retrieval improved visuospatial memory. Similarly, Kang (2010) examined the mnemonic benefits of covertly retrieving in a situation where an overt response would be difficult or time consuming. In Kang's experiments, subjects learned a set of Chinese characters and their English translations. Then subjects either practiced covert retrieval of the Chinese characters by forming a mental image of the characters in response to the English form, or they restudied the pairs across two blocks. On the final retention test, subjects were provided with the English words and were required to draw the Chinese characters. Across three experiments, Kang showed that covertly retrieving the Chinese characters resulted in superior final performance relative to restudying. Finally, Orlando and Hayward (1978) examined the effects of mentally rehearsing text material and found that mental rehearsal improved memory later relative to rereading or taking notes.

These experiments show that covert retrieval benefits retention (Carpenter & Pashler, 2007; Kang, 2010; Orlando & Hayward, 1978; see too Izawa, 1976, for related research). However, these experiments did not address the issue of whether covert retrieval benefits retention to the same degree as overt retrieval. A few experiments are relevant to this issue. Covert and overt responding have been examined in studying the effects of adjunct questions in learning from texts; adjunct questions are those embedded into text materials (see Anderson & Biddle, 1975) and answering such questions can facilitate comprehension and retention. Answering adjunct questions is often not considered retrieval practice as in the testing effect literature because often the subjects have access to the materials while answering the questions. Still, Roediger and Karpicke (2006a) argued that the procedures are rather similar; answering adjunct questions is rather like taking an open-book test (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). However, the adjunct questions literature is mixed with regard to overt and covert responding. Some research has indicated that overtly responding to the adjunct questions results in greater comprehension and retention relative to covertly responding, whereas other research has shown no differences between the two types of responses (e.g., Kemp & Holland, 1966; Michael & Maccoby, 1953). The literature is likely mixed because not all of these experiments ensured that subjects were actually answering the questions covertly in the covert conditions. If subjects are not engaged during covert responding, then one cannot fairly compare overt and covert responding. Other research from the literature on motor skill learning has shown that covert or mental practice can be beneficial as well; for example, Wohldmann, Healy, and Bourne (2008) showed that mental practice produced repetition priming and transfer in a typing task and that under some circumstances mental practice can be advantageous relative to physical practice. However, other researchers (using different tasks) have found that mental practice leads to less improvement than physical practice (Kohl & Roenker, 1983) or, in another case, to no benefit at all (Shanks & Cameron, 2000). These mixed effects are difficult to interpret and principles of mental practice in motor skill tasks may differ considerably than for those in verbal and visuospatial retrieval tasks.

One other recent article deserves mention here, that of Putnam and Roediger (2013). The experiments they reported, conducted in the same lab during the same period as those in the current article, were focused on the issue of type of responding on a first test and whether this manipulation would affect the magnitude of the testing effect. Putnam and Roediger (2013) asked if different forms of overt responding on a first test (spoken, typed) would have different effects on a final test (also spoken or typed). The answer to these questions across several experiments was no, and the type of final test did not matter either. Their experiments also had a covert retrieval condition and so are relevant to the present findings. We consider these results in greater detail in the General Discussion. The Putnam and Roediger (2013) experiments all used paired-associate learning, whereas the current experiments used free recall. If production leads to distinctive processing (relative to covert retrieval), then free recall should provide a more sensitive test.

The purpose of the experiments reported here was to directly test whether an overt response during retrieval produces a superior benefit on later retention relative to covert retrieval. In our experiments, we employed better conditions than have typically been used in past research (such as the adjunct questions literature) to ensure that subjects complied with the request to covertly retrieve as instructed. We used four somewhat different methods across the four experiments to provide generality. In Experiments 1–3 we used a within-subject design to compare overt retrieval, covert retrieval, and a no-test control condition. We used categorized word lists, and the categories used contained differing numbers of items (four, five, and six) so that when subjects were asked to covertly retrieve and report the number of items they recalled they would be less likely to rely on category size for responding. If all of the categories presented the same number if items, then it would be possible for subjects to simply report the maximum number of items possible during covert retrieval. By varying the number of items in categories, we intended to minimize or eliminate this strategy. We also employed free recall as the final assessment measure because we thought this task was most likely to reveal differences between overt and covert retrieval practice, if they exist. In the first experiment, we intermixed overt and covert retrieval trials so that subjects did not know on any trial whether overt retrieval would be required; we hoped to encourage subjects to covertly retrieve when asked to do so and to hold the answer in mind if overt retrieval was requested. In the second experiment, overt and covert retrieval trials were blocked. In the third experiment, we examined overt and covert retrieval practice after a relatively long-term delay (2 days). Finally, in Experiment 4 we compared overt and covert retrieval practice in a between-subjects design, and we included a restudy control condition to determine whether forms of retrieval practice produced testing effects against this more conservative baseline.

Experiment 1

Experiment 1 examined whether covert and overt retrieval produced equivalent retrieval practice effects. Subjects first studied a categorized list and then took an initial recall test in which they were given category names (e.g., *vegetables*) as cues to recall studied words. Categorized lists were used because they naturally afford relational processing (Hunt & Einstein, 1981). If overt responding produces a retention benefit relative to covert retrieval, it is likely due to enhanced item-specific information that is distinctive (as in the production effect). Because categorized lists afford relational processing, we should see a benefit of distinctive processing in overt retrieval relative to covert retrieval (if there is one to be seen) using this paradigm. During the initial test, for some categories, subjects were cued to recall and type the words they recalled (the overt retrieval condition), while for other categories the subjects were cued to recall words they had studied, but they did not type the words (the covert retrieval condition). All subjects were cued to recall for 40 s, and when that time had elapsed, subjects were directed to type the words they had recalled (overt) or not to do so (covert). Therefore, when subjects were given a category cue, they did not know until 40 s later whether they would need to produce the items. This procedure ensured as much as possible that students would initially engage in covert retrieval for the 40 s period. A third set of categories was not cued during the initial test (the no test control condition). After a 15-min delay, subjects completed a final free recall test to assess retention of the items. Again, free recall was used as the final assessment because we thought it best suited for measuring effects of distinctive processing that may be produced by overt responding.

Method

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

Design. Three within-subject conditions were employed: overt retrieval, covert retrieval, and no test. The overt and covert trials were intermixed, as described below. For the no test condition, subjects were not cued to recall items at all during the initial test.

Materials. Materials consisted of categorized word lists. Items were taken from 18 categories from the updated version of the Battig and Montague (1969) word norms (Van Overschelde, Rawson, & Dunlosky, 2004). Six items were drawn from each category. The first four items from each category were not used to help reduce the influence of guessing on the tests (Tulving & Pearlstone, 1966). The categories were divided into three sets of six categories, and each set was fully counterbalanced across the three conditions (overt retrieval, covert retrieval or no test). Subjects studied a list of 90 words, 18 categories with four, five, or six words per category. For each subject, the computer program randomly selected six categories for the four-word condition, six categories for the five-word condition, and six categories for the six-word condition, making certain that each of the three sets of categories contained two categories with four words, two with five words, and two with six words so that the total number of words assigned to each condition was equated. For the four- and five-word conditions, the computer randomly selected four or five words from the six words.

Procedure. At the beginning of the experiment, subjects studied the categorized list of 90 words. The list was blocked by category. Subjects saw a category name for 2 s (e.g., *VEGETABLES*) followed by each word for 2 s (e.g., *cucumber*) with a 500-ms interstimulus interval between words. The order of categories in a list was randomized. In addition, categories assigned to each of the three sets were evenly distributed throughout the study phase such that categories assigned to any of the three conditions did not occur more frequently near the beginning or end of the study list. The items within each category were also randomly ordered for each subject. Subjects were instructed to study the words as they appeared so that they would be able to recall them later.

After the study phase, all subjects played a video game (Pac-Man) on the computer for 3 min. After this filler task, subjects completed the initial test. Before this test, subjects were warned against guessing and were told that the experimenter might ask them to recall the items again later in the experiment. Overt and covert retrieval trials were intermixed during the initial test. Subjects were given the category name and instructed to mentally recall the words they had studied that belonged to the category for 40 s. After 40 s had elapsed, an instruction appeared at the top of the screen. In the overt retrieval condition, a text field appeared, and the subjects were told to type as many of the words as they could recall for 20 s, whereas in the covert retrieval condition, the subjects were told to continue thinking of the words they had recalled, but they were not instructed to type them. At the end of each recall trial subjects typed the total number of words they had recalled from that category (e.g., 3). This was done to gain some estimate of the number of words recalled during covert retrieval.

After subjects completed the initial test, they played a video game (Tetris) on the computer for 15 min. Then, all subjects completed a final free recall test. Subjects were asked to recall words for 10 min by typing as many studied items from as many categories as possible, but they were also warned against guessing. At the end of the experiment the subjects were debriefed and thanked for their participation.

Results

All results were significant at the .05 level, unless otherwise noted. The Bonferroni correction for multiple comparisons was used for all pairwise comparisons.

On the initial test, subjects reported recalling, on average, the same number of items during overt trials (M = 2.93 or 58%, mean correct recall was 2.50) as they did during covert trials (M = 2.94 or 58%; $F < 1$).

The critical data from the final free recall test are provided in Figure 1, which shows about a 20% advantage of practicing both prior overt and covert retrieval, but no difference between the two retrieval practice conditions. A one-way analysis of variance (ANOVA) indicated that there was a difference in final recall among the conditions, $F(2, 70) = 27.40$, $\eta_p^2 = .44$, and pairwise comparisons showed no difference between proportion of items recalled from the overtly retrieved categories (M = .45) and the covertly retrieved categories (M = .47; $F < 1$). However practicing overt retrieval, $F(1, 35) = 32.38$, $\eta_p^2 = .48$, and covert retrieval, $F(1, 35) = 61.29$, $\eta_p^2 = .64$, produced greater final recall than when no retrieval was practiced (the no test condition, M = .26).

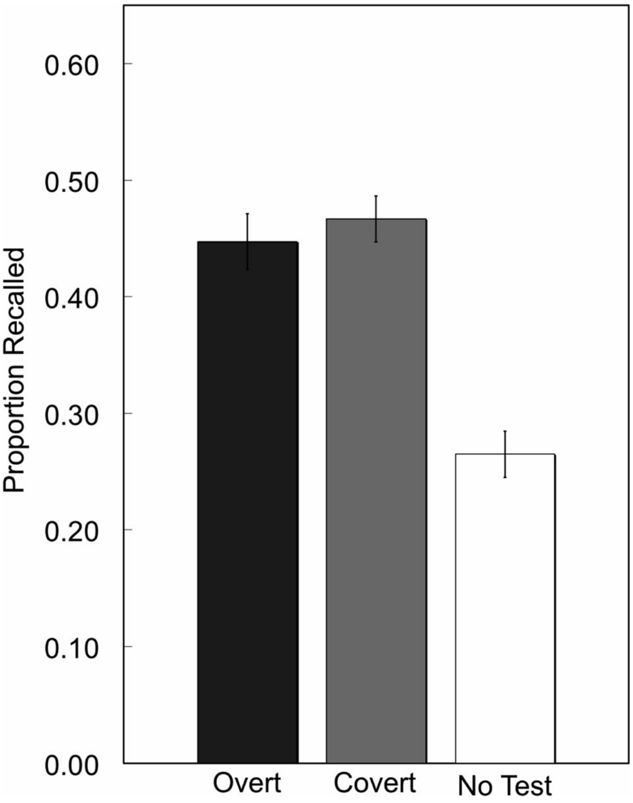

Figure 1. Performance on the final free recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 1. Error bars represent within-subject standard errors of the mean (Cousineau, 2005; Morey, 2008).

The final free recall data were also analyzed in terms of the number of categories recalled (using the convention of crediting category recall if subjects recalled at least one word from the category; Cohen, 1963) and the number of words-per-category recalled (Tulving & Pearlstone, 1966). These data are shown in the top panel of Table 1. Practicing retrieval enhanced the number of categories recalled on the final test but did not affect the number of words recalled within each category. A one-way ANOVA indicated differences in category recall, $F(2, 70) = 44.86$, $\eta_p^2 = .56$; subjects recalled more categories from the overt (M = 4.42) and covert retrieval conditions (M = 4.69) than from the no test condition (M = 2.58), but no significant difference was found between the two retrieval conditions. In addition, no differences were obtained among conditions on the words-per-category recall measure, $F(2, 70) = 1.95$, *ns*.

Table 1
Measures of Category Recall and Words-Per-Category Recall on the Final Free Recall Test in Experiment 1 and 2

Variable	Category recall	Words-per-category	Total recall
Experiment 1			
Overt	4.42 (0.17)	2.95 (0.14)	13.42 (0.72)
Covert	4.69 (0.16)	2.89 (0.11)	14.00 (0.59)
No Test	2.58 (0.18)	2.57 (0.18)	7.94 (0.60)
Experiment 2			
Overt	4.64 (0.15)	2.83 (0.10)	13.69 (0.59)
Covert	4.36 (0.18)	2.90 (0.12)	13.08 (0.67)
No Test	2.69 (0.19)	2.78 (0.13)	8.03 (0.61)

Note. Within-subject standard errors are reported in parentheses (Cousineau, 2005; Morey, 2008). Category recall multiplied by words-per-category recall does not perfectly equal total recall due to rounding.

Discussion

The results of Experiment 1 showed that covert retrieval produced a comparable testing effect relative to overt retrieval. The two conditions produced comparable effects in overall free recall performance and in category access and words-per-category recalled. Further, this experiment created conditions in which the task requirements during initial retrieval were tightly controlled with subjects in both tested conditions engaging in covert retrieval for a 40-s period. This control helped ensure that subjects were complying with instructions in the covert condition, but it may have introduced some artificiality into the retrieval process in the overt condition. Under standard retrieval instructions, subjects usually bring the information to mind (i.e., they have a recollective experience) and then produce an overt response rather quickly thereafter (i.e., memory performance, see Tulving, 1983, pp. 134–137). However, in the overt retrieval condition of Experiment 1, we artificially forced subjects to covertly retrieve category members for a block of time before they produced overt responses. In Experiment 2, we asked whether allowing subjects to retrieve more naturally (i.e., recall the words and immediately report them) in the overt retrieval condition would result in a larger retrieval practice effect for the overt relative to the covert retrieval condition.

Experiment 2

Experiment 2 employed the same three within-subject conditions to address whether covert and overt retrieval produced equivalent retrieval practice effects. However, instead of intermixing the overt and covert retrieval trials as in Experiment 1, subjects completed two blocks of initial tests, one for overt retrieval and one for covert retrieval. During one initial test, subjects were cued with category names and were instructed to overtly retrieve the items as well and as quickly as possible. Unlike the procedure in Experiment 1, subjects in Experiment 2 were permitted to type the words as they came to mind. During the covert test subjects were instructed to bring items to mind but not to type them. A third set of categories was not cued during the initial test (the no test control condition).

Method

Subjects, materials, and design. Thirty-six subjects (20 female, ages 18–30 years, median age of 20) were recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or pay. None had participated in Experiment 1. Two subjects were replaced because they did not follow testing instructions. The materials and design were identical to those used in Experiment 1.

Procedure. The procedure for Experiment 2 was similar to that of Experiment 1, with two differences: First, covert and overt retrieval trials were separated into two blocks (order was counterbalanced across subjects). During each block, subjects were presented with the category names assigned to the appropriate retrieval condition one at a time for 60 s. Second, during overt retrieval subjects were permitted to type the words into the computer as they came to mind, and during covert retrieval subjects typed an X for each word they recalled as they came to mind. Importantly, they never typed the specific items during the covert trials. In all other respects the procedure was identical to the one used in Experiment 1.

Results

As in Experiment 1, the numbers of words recalled or reported on the initial test were nearly identical in the overt and covert retrieval conditions. Subjects produced, on average, the same number of items per category during the overt test (writing out the words, $M = 3.17$ or 63%, mean correct recall was 2.67) and the covert test (indicated by entering X for items recalled, $M = 3.21$ or 64%; $F < 1$). The same pattern of results was obtained regardless of the order in which the initial tests were taken.

The results from the final free recall test are shown in Figure 2 and show the same pattern as in Experiment 1, with robust effects of retrieval practice from prior covert and overt retrieval conditions and essentially no difference between these conditions. A one-way ANOVA indicated that there were differences among the initial conditions, $F(2, 70) = 24.79$, $\eta_p^2 = .42$. Pairwise comparisons showed that free recall of items from the overt ($M = .46$) and the covert ($M = .44$; $F < 1$) retrieval conditions were not significantly different from one another. However practicing overt retrieval, $F(1, 35) = 48.63$, $\eta_p^2 = .58$, and covert retrieval, $F(1, 35) = 29.33$, $\eta_p^2 = .46$, produced greater final recall than when no retrieval was practiced (the no test condition, $M = .27$). This pattern of results was the same regardless of the order in which the initial tests were taken.

Category recall and words-per-category recall from Experiment 2 are shown in the bottom panel of Table 1 and again show the same pattern as in Experiment 1. Practicing retrieval enhanced the number of categories recalled on the final test but did not affect the number of words recalled within each category. A one-way ANOVA showed differences in category recall, $F(2, 70) = 36.18$, $\eta_p^2 = .51$. Subjects recalled more categories from the overt ($M = 4.64$) and covert retrieval ($M = 4.36$) conditions than from the no test condition ($M = 2.69$), but there was no difference between the two retrieval conditions. No differences were obtained among conditions on the words-per-category recall measure ($F < 1$).

Discussion

In Experiment 2, we replicated the results from Experiment 1 using a procedure that allowed subjects to retrieve more naturally

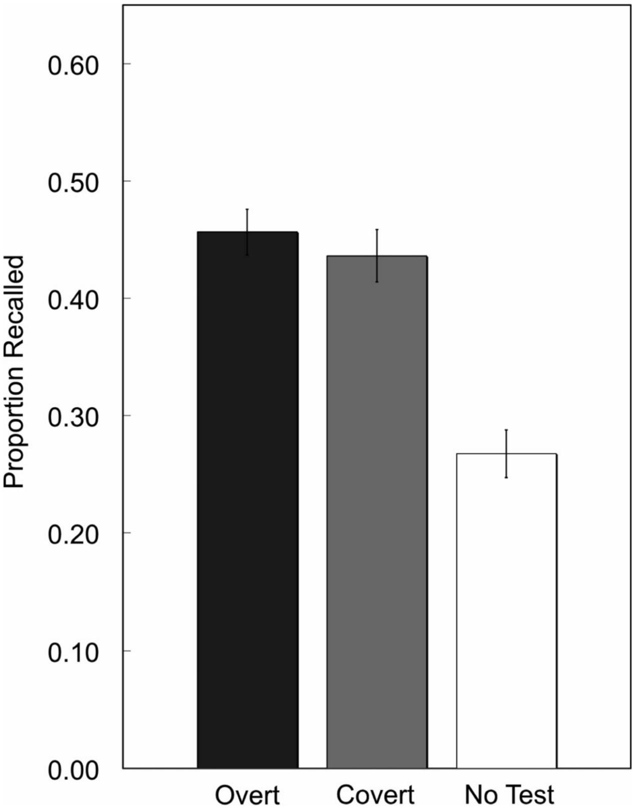

Figure 2. Performance on the final free recall test for the overt retrieval, covert retrieval, and no test conditions in Experiment 2. Error bars represent within-subject standard errors of the mean (Cousineau, 2005; Morey, 2008).

in the overt retrieval condition. Retrieval practice effects of comparable magnitude were obtained for both overt and covert retrieval conditions. Once again, practicing retrieval enhanced category recall on the final test, relative to the no test control condition, but there were no differences in recall of words-per-category, indicating that both overt and covert retrieval improved subjects' ability to access the categories but did not affect the number of words recalled once a category was accessed.

Experiment 3

In both Experiments 1 and 2, retention was measured after a short delay (15 min). Because the effects of retrieval practice sometimes change over time (e.g., Hogan & Kintsch, 1971), we thought it important to determine whether differences between the two retrieval conditions would arise after a long delay. In Experiment 3 we compared covert retrieval and overt retrieval after both a short (15-min) and long (2-day) delay to ask whether the two conditions would still provide comparable retention benefits. Accordingly, in Experiment 3 one group of subjects completed the final free recall test during the initial learning session as in the first two experiments, whereas another group of subjects completed the final retention test after a 2-day delay.

Method

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

Materials. Sixteen of the categories from the first two experiments were used, and similar to the earlier experiments either five or six items were presented per category. When five items were studied, the computer randomly determined which of the six items were presented to each subject. The categories were divided into four sets of four categories, one set for each of the four conditions.

Design. Experiment 3 used a 4 (learning condition) × 2 (retention interval) mixed factorial design, with learning condition manipulated within-subject and retention interval manipulated between subjects. There were four learning conditions: two retrieval practice conditions (overt retrieval and covert retrieval) and two control conditions, restudy and no test. The overt retrieval, covert retrieval, and no test conditions were the same as in Experiment 2. During the restudy condition, subjects were presented with the items in each category assigned to the restudy set one at a time. As in Experiment 2, the conditions were blocked during the initial phase. The order of the category sets was held constant, and the initial test conditions were fully counterbalanced across the sets. Retention interval was manipulated between subjects; some subjects completed the final tests 15 min after the learning phase (the immediate condition), and the other group returned 2 days later to complete the final tests (the delayed condition).

Procedure. The procedure for Experiment 3 was similar to that of Experiment 2, with three differences. First, a restudy condition was added. During the restudy phase, subjects restudied items from the four categories assigned to the restudy condition. As in the first study phase, the category name was presented first for 2 s in all uppercase letters followed by the items in each category. However, the interstimulus interval was lengthened by 3,750 ms so that the restudy phase took the same amount of time as the overt and covert initial tests. Second, subjects were given 30 s to retrieve during each retrieval trial. This was done so that the testing conditions and the restudy condition could be equated for time and because subjects in the first two experiments reported that they had more time than was necessary to retrieve for each category cue. Finally, subjects in the immediate retention condition completed the final free recall test during the first session of the experiment 15 min after the study phase. Subjects in the delayed retention condition returned to the lab 2 days later to complete the final free recall test. In all other respects the procedure was identical to the one used in Experiment 1.

Results

The number of items produced either overtly or covertly during the initial tests was nearly identical. Collapsed across retention interval, an average of 3.24 items (59%) were produced in the overt condition (2.45 correct) and 3.06 (56%) in the covert condition (indicating by entering X for items recalled). There was no significant difference between the two ($F = 1.38$).

The results from the final free recall test are shown in Table 2. On the immediate final test, the effect of retrieval practice obtained

Table 2
Proportion Correct on the Final Free Recall Test and Forgetting Across the Delay for the Overt Retrieval, Covert Retrieval, Restudy, and No Test Conditions in Experiment 3

Variable	Overt	Covert	Restudy	No Test
Immediate	.34 (.03)	.32 (.03)	.58 (.05)	.17 (.03)
Delayed	.27 (.03)	.25 (.03)	.29 (.03)	.07 (.03)
Proportional forgetting	.21	.21	.49	.59

Note. Within-subject standard errors are reported in parentheses where applicable (Cousineau, 2005; Morey, 2008).

in Experiments 1 and 2 relative to the no test condition was replicated. Restudying produced much higher performance on this test, which is no surprise because subjects in the retrieval practice conditions reexperienced only what they could successfully recall (about 45% of the items) while subjects in the restudy condition reexperienced 100% of the list. On the 2-day delayed test, all three reexposure conditions (covert test, overt test, and restudy) showed superior performance to the no test condition. However, the initial test produced a dramatic drop in proportional forgetting relative to the restudy and no test groups.

A 4 (initial test condition) × 2 (retention interval) ANOVA with repeated measures on the first factor revealed that overall there were differences among the initial test conditions, $F(3, 138) = 29.46$, $\eta_p^2 = .39$, and forgetting occurred overall—subjects in the immediate group ($M = .35$) performed significantly better than those in the delayed group ($M = .22$), $F(1, 46) = 8.74$, $\eta_p^2 = .16$. However, these effects were qualified by a significant interaction, $F(3, 138) = 4.71$, $\eta_p^2 = .09$. The interaction revealed that restudying resulted in superior short-term retention relative to retrieving the items (either overtly or covertly) or doing nothing (the no test condition), but this advantage did not hold after a longer delay (see Roediger & Karpicke, 2006b, for a similar pattern).

Post hoc analyses confirmed these observations. Subjects in the immediate test group recalled significantly more items from the restudied categories ($M = .58$) than from categories overtly recalled ($M = .34$), covertly recalled ($M = .32$) and those not tested ($M = .17$). In addition, these subjects recalled significantly fewer items from the nontested categories relative to items from categories assigned to the other three conditions. Importantly, recall from the overtly tested categories and the covertly tested categories did not differ. A slightly different pattern of results was found for subjects in the delayed test condition. For these subjects, recall of items from the nontested categories ($M = .07$) was significantly worse than recall from the overtly tested categories ($M = .27$), covertly tested categories ($M = .25$), and restudied categories ($M = .29$). No other comparisons reached significance. It is highly likely that practicing retrieval did not produce better retention than restudying after the delay because of differences in item reexposure. The data suggest that overt or covert retrieval practice may result in less forgetting than restudying or not taking an initial test. Proportional measures of forgetting, (initial recall − final recall)/initial recall (see Roediger & Karpicke, 2006b), indicated that both overt or covert retrieval practice resulted in only 21% forgetting. Forgetting following restudy of the category members or not taking an initial test resulted in much more forgetting than practicing retrieval did (49% and 59% forgetting, respectively). Had we matched initial retrieval success (e.g., by bringing subjects up to criterion; Karpicke & Roediger, 2008; Karpicke & Smith, 2012), we would probably have seen an advantage of retrieval practice over restudying on a delayed retention test because that is the typical outcome in the literature.

Category recall and words-per-category recall from Experiment 3 are shown in Table 3, revealing somewhat different patterns on immediate and delayed tests. A 4 (initial test condition) × 2 (retention interval) ANOVA on the category recall results revealed that overall there were differences among the conditions, $F(3, 138) = 30.29$, $\eta_p^2 = .40$. Post hoc comparisons indicated that category recall was significantly lower for categories that were not tested ($M = 0.94$) relative to categories that were overtly tested ($M = 2.19$), covertly tested ($M = 2.10$), and restudied ($M = 2.71$). In addition, category recall was higher for categories that were restudied ($M = 2.71$) relative to those that were covertly retrieved ($M = 2.10$), although this outcome mainly occurred because of high category recall on the same-day test. No other comparisons among the four conditions reached significance. There was also a main effect of retention interval, $F(1, 46) = 5.26$, $\eta_p^2 = .10$, indicating that forgetting occurred from the immediate to the delayed final tests. Subjects in the immediate group ($M = 2.26$) recalled significantly more categories than those in the delayed group ($M = 1.71$). The interaction only reached a marginal level of significance, $F(3, 138) = 2.14$, $p = .10$, $\eta_p^2 = .04$. Most important, as in the previous two experiments, category recall between the overtly and covertly tested categories did not differ.

Differences were also obtained in the words-per-category recall measure. In Experiment 3, words-per-category recall showed the same results as category recall. A 4 (initial test condition) × 2 (retention interval) ANOVA on the words-per-category recall results revealed a significant main effect of initial condition, $F(3, 138) = 14.94$, $\eta_p^2 = .25$. Post hoc comparisons indicated that overall words-per-category recall was significantly lower for categories that were not tested ($M = 1.59$) relative to categories that were overtly tested ($M = 2.50$), covertly tested ($M = 2.40$), and restudied ($M = 3.13$). In addition, words-per-category recall was higher for categories that were restudied ($M = 3.13$) relative to those that were covertly retrieved ($M = 2.40$); again, this difference was mainly due to high item recall in the same-day test

Table 3
Measures of Category Recall and Words-Per-Category Recall on the Final Free Recall Test in Experiment 3

Variable	Category recall	Words-per-category	Total recall
Immediate condition			
Overt	2.33 (0.19)	2.73 (0.26)	7.50 (0.03)
Covert	2.21 (0.17)	2.53 (0.19)	7.00 (0.03)
Restudy	3.25 (0.24)	3.82 (0.27)	12.71 (0.05)
No Test	1.25 (0.16)	2.06 (0.26)	3.79 (0.03)
Delayed condition			
Overt	2.04 (0.18)	2.26 (0.20)	5.96 (0.03)
Covert	2.00 (0.19)	2.27 (0.17)	5.50 (0.03)
Restudy	2.17 (0.18)	2.44 (0.21)	6.46 (0.03)
No Test	0.63 (0.21)	1.13 (0.26)	1.54 (0.03)

Note. Within-subject standard errors are reported in parentheses (Cousineau, 2005; Morey, 2008). Category recall multiplied by words-per-category recall does not perfectly equal total recall due to rounding.

condition. No other comparisons reached significance. There was also a main effect of retention interval, $F(1, 46) = 6.20$, $\eta_p^2 = .12$, showing forgetting over the 2 days. Subjects in the immediate group ($M = 2.79$) recalled significantly more words-per-category than those in the delayed group ($M = 2.02$). The interaction only reached a marginal level of significance, $F(3, 138) = 2.31$, $p = .08$, $\eta_p^2 = .05$. Most important, as in the previous two experiments, words-per-category recall between the overtly and covertly tested categories did not differ.

Discussion

In Experiment 3, we replicated our previous results on a final free recall test completed 15 min after learning showing that covert retrieval practice produces the same benefit as overt retrieval practice. Importantly, we showed that the same outcome occurred after a 2-day retention interval. Moreover, analyses of category recall and items-per-category were the same on the immediate and delayed retention tests.

Experiment 4

As noted above, one can only expect retrieval practice to outperform restudying items when retrieval practice is successful. If performance during retrieval practice is not very high, then the greater exposure from the restudy condition will overwhelm the retrieval practice benefit (Karpicke & Bauernschmidt, 2011; Karpicke & Smith, 2012). The purpose of Experiment 4 was to examine overt and covert retrieval under conditions designed to increase levels of initial retrieval success. During retrieval practice, we provided more powerful cues to boost performance. This allowed us to compare retrieval practice to restudying and also to see if overt and covert retrieval practice would still produce equivalent benefits with higher levels of retrieval success. An additional feature was to include no response during covert retrieval. In the first three experiments covert retrieval still involved some type of response (typing in the number of items in Experiment 1, typing an "X" in Experiments 2 and 3). In Experiment 4 we removed all forms of overt responding from the covert condition. Finally, overt and covert responding were contrasted in between subjects comparisons rather than the within-subject comparisons of the previous experiments.

Method

Subjects. Sixty subjects (38 female, ages 18–44 years, median age of 20) were recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or pay. None of the subjects had participated in Experiments 1, 2, or 3.

Design. Two within-subject conditions were employed: retrieval and restudy. In addition, retrieval format was manipulated between subjects. Half of the subjects were instructed to practice overt retrieval (they typed the item during the initial test), and half of the subjects were instructed to practice covert retrieval (they thought of the item during the initial test but did not produce it). For the restudy condition, subjects read the words on the screen.

Materials. Materials consisted of categorized word lists using 10 categories from the previous experiments. Six items from each category were used for a total of 60 items. Five categories were assigned to the restudy condition and five categories were assigned to the retrieval condition. The assignment was counterbalanced such that each set of five categories was assigned to each of the two conditions an equal number of times across subjects.

Procedure. The experiment began with a study phase during which subjects saw pairs of category names and items for 2 s each followed by a 500-ms interstimulus interval. The name of a category was presented in all capital letters (e.g., *VEGETABLES*), and the item from the category (e.g., *cucumber*) was shown below the category name. Whereas the study list was blocked by category in the three previous experiments, in Experiment 4 the order of words within the list was randomized. After the study phase, all subjects completed a filler task (playing Pac-Man) for 3 min, and after the filler task subjects completed the initial test. During the initial test, retrieval and restudy trials were intermixed, and each trial lasted for 6 s. During retrieval trials, subjects saw a category name and the first two letters of the target word (e.g., *VEGETABLES – cu*_____) and were instructed to recall the word that they had studied that completed the word stem. Importantly, subjects were instructed to think back to the original study list in order to complete the word stem (see Karpicke & Zaromb, 2010). Subjects assigned to practice overt retrieval typed the word into a text box. They were instructed to type the full word including the first two letters that were already provided. Subjects assigned to practice covert retrieval thought of the correct item but did not make any type of physical response. During restudy trials, the word was presented intact below the category name, and subjects were instructed to silently study the word. After subjects completed the initial test, they played Tetris for 15 min and then were instructed to recall all target words from the experiment in any order.

Results

During initial retrieval, subjects in the overt retrieval condition produced the correct item when cued with the category name and the first two letters of the item 72% of the time. Because subjects in the covert condition did not make any type of response during the initial test, we cannot report how many items they correctly recalled, but our instructions emphasized that they should attempt covert retrieval.

The critical results from the final free recall test are shown in Figure 3. There was a large effect of practicing retrieval over studying for subjects in both the overt and covert retrieval conditions, despite the fact that in the restudy condition subjects were exposed to 100% of the items a second time relative to only 72% recall in the overt retrieval condition. A 2 (retrieval vs. restudy) × 2 (overt vs. covert) ANOVA with repeated measures on the first factor revealed that there was an advantage of practicing retrieval over restudying, $F(1, 58) = 27.17$, $\eta_p^2 = .32$. There was also an advantage of covert retrieval over overt retrieval, $F(1, 58) = 5.08$, $\eta_p^2 = .08$. There was no interaction ($F < 1$). The key finding in Experiment 4 was that both overt retrieval practice, $F(1, 29) = 9.26$, $\eta_p^2 = .24$ (*M*s = .53 vs. .43) and covert retrieval practice, $F(1, 29) = 19.79$, $\eta_p^2 = .41$ (*M*s = .63 vs. .51) produced advantages relative to the restudy control conditions, with the size of the retrieval practice effects being roughly equivalent in the overt and covert retrieval conditions.

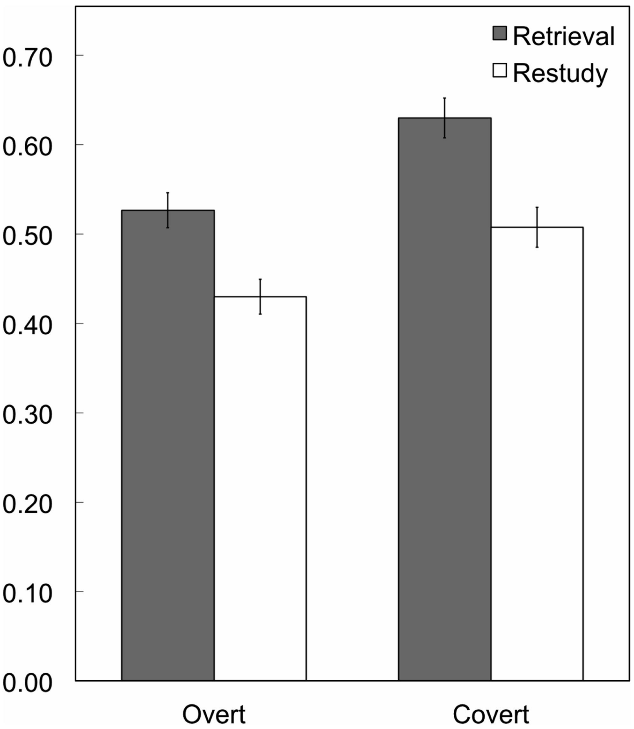

Figure 3. Performance on the final free recall test for the overt retrieval and covert retrieval conditions, and from the restudy controls in Experiment 4. Error bars represent within-subject standard errors of the mean (Cousineau, 2005; Morey, 2008).

Category recall and words-per-category recall from Experiment 4 are shown in Table 4. A 2 (retrieval vs. restudy) × 2 (overt vs. covert) ANOVA with repeated measures on the first factor showed that practicing retrieval led to greater category recall than restudying, $F(1, 58) = 17.38$, $\eta_p^2 = .23$. Those in the covert retrieval condition performed marginally better than those in the overt retrieval condition, $F(1, 58) = 3.10$, $p = .08$, $\eta_p^2 = .05$. There was no interaction ($F < 1$). In Experiment 4 there were also differences for words-per-category recall. A 2 (retrieval vs. restudy) × 2 (overt vs. covert) ANOVA with repeated measures on the first factor showed that practicing retrieval resulted in greater words-per-category recall than restudying, $F(1, 58) = 25.72$, $\eta_p^2 = .31$. As with the other measures, those in the covert retrieval condition performed better than those in the overt retrieval condition, $F(1, 58) = 5.29$, $\eta_p^2 = .08$. There was no interaction ($F < 1$).

Discussion

In Experiment 4, we showed that when retrieval success is boosted, both overt and covert retrieval practice benefit memory more than restudying the items. This was true even though restudying the items still potentially provided an advantage; in the restudy condition all subjects were reexposed to 100% of the items, whereas in the overt retrieval practice conditions subjects produced 72% of items. In this experiment, subjects in the covert retrieval condition performed better on the retention test than those in the overt retrieval conditions. However, in light of the absence of an interaction, it is plausible that this came about due to assignment of subjects to each of the two between-subjects conditions. Note that subjects in the restudy condition combined with covert retrieval performed better than those in the restudy condition combined with the overt retrieval condition, despite the fact that the study trials were exactly the same in the two conditions. It is also possible that covert retrieval practice truly produces a recall advantage relative to overt retrieval under certain circumstances. We discuss these possibilities in the General Discussion.

General Discussion

The results of all four experiments converge on the conclusion that covert retrieval practice provides as much of a benefit as overt retrieval practice on a later test of retention, at least with the materials and procedures used here. In the first two experiments, practicing retrieval on an initial test resulted in superior recall of the categorized word lists on a later test relative to a no test control, but no difference was obtained between covert and overt retrieval under either tightly controlled (Experiment 1) or more natural (Experiment 2) retrieval conditions. In Experiment 3, we replicated these results on a final free recall test completed 15 min after learning, and we also showed that the same outcome occurred after a 2-day retention interval. In Experiment 4, practicing overt and covert retrieval both resulted in better performance on a later test relative to a restudy control. Experiment 4 also demonstrated that covert retrieval produces at least as much of a retrieval practice benefit relative to overt retrieval even when the form of retrieval is manipulated between subjects. Taken together, this set of experiments provides evidence that retrieval practice improves retention, but does not provide support for the notion that overt and covert retrieval practice produce differential effects on later memory. In one experiment in the set (Experiment 4), covert retrieval produced a slightly greater effect than did overt retrieval. However, students in the covert retrieval condition performed better on the study items than those in the overt retrieval condition. This result could potentially lend support to the idea mentioned previously: The differences between overt and covert retrieval practice could be due to random assignment of subjects to conditions. It is possible that the performance differences in the restudy conditions came about because the lists were mixed. The different types of retrieval practice within each list (overt or covert) could have had an effect on the restudy items mixed within the list, creating a difference between the restudy items as well (see Greene, 1989). However, given that an advantage of covert retrieval practice was

Table 4
Measures of Category Recall and Words-Per-Category Recall on the Final Free Recall Test in Experiment 4

Variable	Category recall	Words-per-category	Total recall
Overt condition			
Retrieval	4.57 (0.16)	3.13 (0.14)	15.80 (0.67)
Study	3.97 (0.16)	2.58 (0.14)	12.90 (0.67)
Covert condition			
Retrieval	4.87 (0.13)	3.78 (0.12)	18.90 (0.58)
Study	4.23 (0.13)	3.05 (0.12)	15.23 (0.58)

Note. Within-subject standard errors are reported in parentheses (Cousineau, 2005; Morey, 2008). Category recall multiplied by words-per-category recall does not perfectly equal total recall due to rounding.

only obtained once, and in absence of an interaction, it seems questionable that covert retrieval truly produces a greater mnemonic benefit than overt retrieval. The results from the four experiments provide support for the idea that practicing retrieval enhances later recall due to the process of bringing a prior experience to mind. Both overt and covert retrieval require this process (Karpicke & Zaromb, 2010; Tulving, 1983, pp. 134–137). Other research described earlier by Putnam and Roediger (2013) provided converging evidence on this point. In their series of paired-associate learning experiments, they found that both speaking and typing responses on a first test led to about the same benefit as covert retrieval on the final criterial test given later. We conclude that covert responding on tests produces as great an enhancement in both cued recall and free recall as does overt responding.

Still, this conclusion is based on accepting the null hypothesis. Therefore, we conducted a mini meta-analysis to provide a quantitative estimate of the actual difference between overt and covert retrieval effects on a later memory test. To do this, we used both the independent comparisons of overt and covert retrieval reported in this article and those reported in Putnam and Roediger (2013). Table 5 shows the 10 comparisons included and the relevant characteristics of the experiments. Each of these experiments was designed to examine relative differences on a second test between overt and covert retrieval practice on a first test, but the various experiments used slightly different methods to do so. Conducting a meta-analysis using both sets of experiments allows us to more precisely estimate the size of the overt and covert retrieval practice effect on later retention than could be done using only the experiments reported here or those reported in Putnam and Roediger (2013).

We first calculated effect sizes (Cohen's *d*) for all comparisons from the raw data, coded such that a positive effect size indicates an advantage of overt over covert retrieval practice on a later test. We then calculated the weighted effect sizes for each independent comparison. The weighted effect size took both the effect size *d* and the power of the design into account. For within-subject comparisons, the correlation between the two measures was taken into account and a within-subject formula for calculating the weight of each within-subject effect size was used (see Ellis, 2010). Because we intended to compare the effect sizes across all comparisons, and slightly different methods were used in each comparison, we applied a random effects model to the data assuming that variability between effect sizes was due to error in sampling and variability in the population of effects. Figure 4 depicts the mean effect sizes and 95% confidence intervals around the effect sizes. Confidence intervals were constructed using ESCI software (Cummings, 2012).

The results of the overall meta-analysis are shown at the very bottom of Figure 4. Using all 10 comparisons of overt and covert retrieval, the meta-analysis estimated the effect size between overt and covert retrieval to be $d = -0.0027$ or zero for all practical purposes. One could argue that using a cue-only delayed judgment of learning (JOL) as a way to induce covert retrieval is not a "pure" manipulation of covert retrieval. For this reason, we also included an estimate of the overt vs. covert retrieval practice effect only including evidence from the studies that used a direct manipulation of covert retrieval (i.e., excluding Experiments 1 and 2 from Putnam & Roediger, 2013). The meta-analysis for direct manipulations of covert retrieval estimated the effect size to be $d = -0.14$, a value still quite close to zero. Looking at the full set of experiments in this article and Putnam and Roediger (2013), it is clear that there is no evidence for a difference between overt and covert retrieval practice on a later memory test.

Figure 4 does shows one study that is potentially different from the others. Experiment 4 from this article produced an effect size estimate of $d = -0.65$, and the 95% confidence interval did not include zero. This result may be due to chance factors, as was discussed earlier. It is also possible that the characteristics of this particular experiment led to a true advantage of covert retrieval practice relative to overt retrieval practice. Why? One possibility is

Table 5
Studies Included in the Meta-Analysis, the Characteristics of the Covert Retrieval Manipulation, the Delay Between Initial Learning and the Final Assessment Test, and the Characteristics of the Final Test

Study	Initial covert retrieval manipulation	Delay	Final test
Putnam & Roediger (2013), E1	Students read the cue only and then made judgments of learning (JOLs)	2 days	Typed cued recall
Putnam & Roediger (2013), E1	Students read the cue only and then made JOLs	2 days	Spoken cued recall
Putnam & Roediger (2013), E2	Students read the cue only and then made JOLs	2 days	Typed cued recall
Putnam & Roediger (2013), E2	Students read the cue only and then made JOLs	2 days	Spoken cued recall
Putnam & Roediger (2013), E3	Subjects brought the target word to mind and then responded as to whether they correctly remembered the word	2 days	Typed cued recall
Smith, Roediger, & Karpicke, E1	Students thought about the words and typed in a number for each category	15 min	Typed free recall
Smith, Roediger, & Karpicke, E2	Students thought of the words and typed an "X" for each word they remembered	15 min	Typed free recall
Smith, Roediger, & Karpicke, E3	Students thought of the words and typed an "X" for each word they remembered	15 min	Typed free recall
Smith, Roediger, & Karpicke, E3	Students thought of the words and typed an "X" for each word they remembered	2 days	Typed free recall
Smith, Roediger, & Karpicke, E4	Students just thought of the correct answer to the category cue and word stem	15 min	Typed free recall

Note. E = experiment; Smith, Roediger, & Karpicke = the current article.

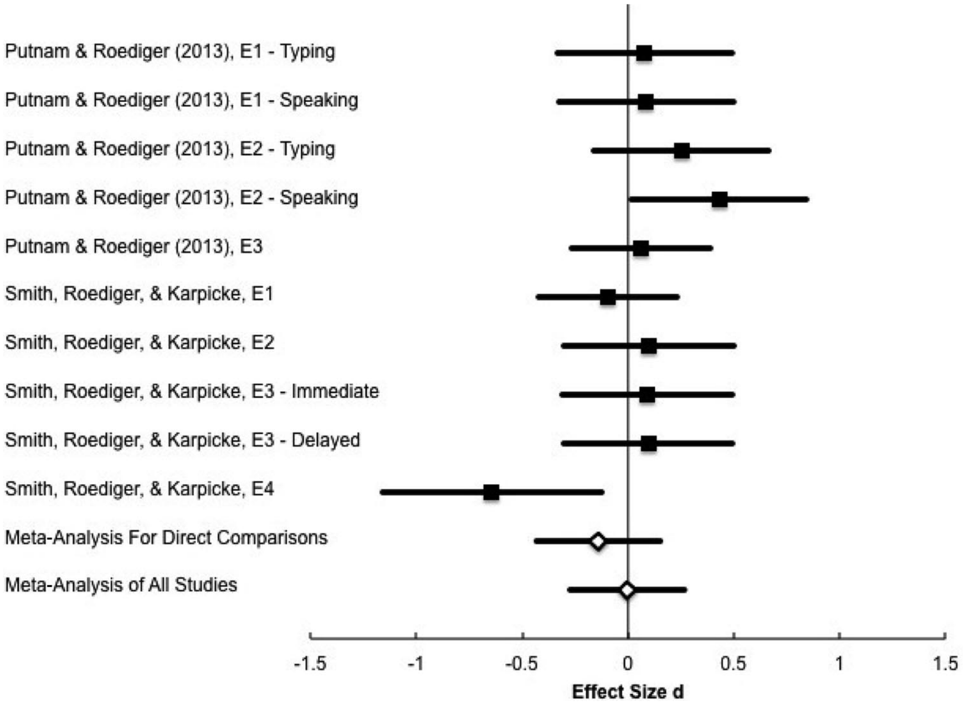

Figure 4. Forest plot depicting effect sizes and 95% confidence intervals around the effect sizes between overt and covert retrieval practice, where a positive effect indicates an advantage for overt over covert retrieval practice.

that this experiment was the only one in this set that did not require *any* overt response during covert retrieval; the other experiments required subjects to make some small overt response (e.g., typing an "X" for every word recalled). It is quite possible that removing all forms of responding during covert retrieval benefits retention in some way. Future research will be necessary in order to determine exactly what conditions of covert retrieval lead to the same learning outcome as overt retrieval, and which conditions, if any, lead to a benefit of covert retrieval over overt retrieval practice. However based on the evidence reported here and in Putnam and Roediger (2013), we do not find strong evidence in favor of a difference between the two forms of retrieval practice and certainly no evidence for an advantage of overt responding.

Despite the fact that the current experiments and those of Putnam and Roediger (2013) do not provide evidence for different effects between overt and covert retrieval on a later test, they do present strong evidence that retrieval practice benefits retention above control conditions. The benefit of retrieval practice on retention is consistent and robust in these experiments and is consistent with the previous literature on the topic. To demonstrate the powerful effect of retrieval practice on later retention, we conducted a mini meta-analysis using the same methods as the previous meta-analysis. Results from the meta-analysis are shown in Figure 5, where a positive effect size (*d*) indicates an advantage of retrieval practice (combining overt and covert retrieval practice) over control conditions (study once control for nine comparisons, and a restudy control for one comparison). Using all 10 comparisons, the meta-analysis estimated the effect size of retrieval practice over control conditions to be $d = 1.10$, which is a large effect size. The effect of practicing retrieval seems to be very large, whereas the overall effect comparing overt and covert retrieval practice is functionally zero.

The fact that performance on the initial covert test cannot be scored may seem like cause for concern. In fact, this is probably the primary reason that covert retrieval is not frequently employed in experiments examining the effects of retrieval practice via testing on later retention. In the first three experiments reported here we attempted to overcome this difficulty by having subjects signal how many items they recalled. Using this technique, we observed no difference in overt and covert retrieval on the initial test when using this indirect measure of covert retrieval. Of course, we have no way of knowing which items were retrieved, but the fact that subjects reported retrieving the same number of items per category in the first three experiments bolsters the assumption that subjects were engaged in similar processes during the two types of initial test. In addition, the procedure in Experiment 1 in which all subjects covertly retrieved (and then only the overt group produced responses) probably equated retrieval for the first 40 s. Subjects were not aware when they began each trial whether they would be required to produce the remembered items until after they spent time covertly retrieving items, thus providing a motivation to covertly retrieve on each trial. The results of Experiment 1, with tight control, were like those of the later experiments with more natural retrieval processes. Therefore, it seems unlikely that subjects retrieved different numbers of items during the overt and covert retrieval attempts.

The present results showing roughly equivalent testing effects from overt and covert responding may seem surprising from some

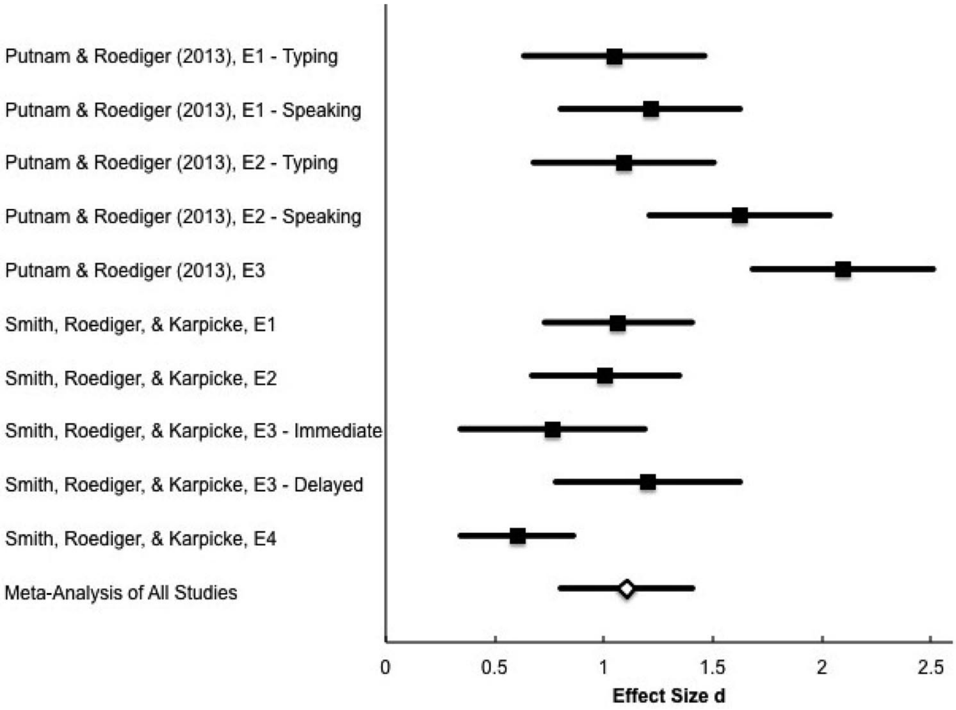

Figure 5. Forest plot depicting effect sizes and 95% confidence intervals around the effect sizes between retrieval practice (overt and covert retrieval practice combined) and control conditions, where a positive effect indicates an advantage for retrieval practice. In all studies, the no test control condition was used when available.

viewpoints. Researchers dating at least back to Robinson (1941) have suggested that overt responding should provide a greater benefit than covert responding because the former involves kinesthetic and other sensory information that might augment performance, in line with more recent research in embodied cognition approaches (e.g., Wilson, 2002). In addition, the production effect experiments of MacLeod et al. (2010) showed that producing items (albeit during a study session, not a test session) relative to silent reading produces a benefit. Nonetheless, results from our four experiments showed no hint of a greater effect of retrieval practice from overt than from covert retrieval.

One criticism of using retrieval practice in educational settings to improve students' retention is that creating and grading tests takes large amounts of time (see Roediger et al., 2011). However, if students create retrieval cues (or questions) for themselves while studying and use covert retrieval to recall the relevant information in later study periods, then retrieval practice (via self testing) is feasible. Even given our results, however, we believe that overt responding during retrieval practice may be desirable in some situations such as the classroom, because unless students believe they may be called upon to produce a response, they may not make the effort to retrieve it. In addition, our results use categorized list paradigms, and we need comparisons of covert and overt retrieval using more complex materials before generalizing too widely.

Practicing retrieval can be beneficial in many other ways that are relevant for education beyond just directly improving retention; Roediger et al. (2011) discussed 10 possible benefits of using testing as a retrieval practice activity. For some of these benefits retrieval practice requiring overt retrieval is needed. For example, when tests are used for formative assessment purposes, to let the instructor know what material students are struggling to understand, then overt responding is required (for the teacher's benefit, if not for the students' benefit). Also, if teachers ask questions to a class and want them to covertly retrieve an answer, having at least one person provide the answer overtly permits other students to know if their tentative answer was correct. Thus, even though overt and covert responding seem to produce equivalent retrieval practice effects under many conditions, overt responding may still be useful in many circumstances. However, when self-testing during study periods, covert responding should be sufficient.

Practicing retrieval can also help students to improve their metacognitive monitoring—how accurate they are at judging how well they know the material—relative to restudying. When students repeatedly restudy their materials they are often overconfident, but practicing retrieval helps to reduce such confidence (e.g., Roediger & Karpicke, 2006b; Karpicke & Roediger, 2008). Retrieval practice can also be used to identify what students know and do not know and can guide further efficient study. In fact, when students use self-testing as a study strategy, it is typically used to exploit this benefit (Karpicke, 2009; Karpicke, Butler, & Roediger, 2009). Of course just as experiments examining the direct effects of practicing retrieval have employed overt responses during retrieval, experiments demonstrating the metacognitive benefits of retrieval practice have required overt responses during retrieval. It is possible that covert retrieval practice may not help students identify gaps in knowledge and improve metacognitive monitoring as well as overt retrieval practice does. Although we know of no direct tests of this idea, in his book about effective study strategies, Robinson (1941) recommended that students recite their lessons overtly rather than covertly for the purposes of diagnosing the state of their knowledge: "[writing out the answer] is more effective since it forces the reader actually to verbalize the answer, whereas a mental review may often fool a reader into believing that a vague feeling of comprehension represents mastery" (p. 30). Further research will be needed to determine whether overt and covert retrieval practice affect students' metacognitions in the same way.

Overall, the results of the four experiments reported here, as well as the meta-analysis conducted using the experiments from this article and from Putnam & Roediger, 2013, indicate that the testing effect produced by covert retrieval practice is just as great as that from overt retrieval. We conclude that covert retrieval is as effective as overt retrieval when care is taken to ensure that students have carried through with covert retrieval, and in some situations may even produce larger gains than overt retrieval practice.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *The Psychological Review: Monograph Supplements, 11,* 159–177. doi:10.1037/h0093018

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. doi:10.1002/acp.1391

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9, pp. 89–132). New York, NY: Academic Press. doi:10.1016/S0079-7421(08)60269-8

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology, 80,* 1–46. doi:10.1037/h0027577

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941–956. doi:10.1002/acp.1239

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14,* 474–478. doi:10.3758/BF03194092

Cohen, B. H. (1963). An investigation of recoding in free recall. *Journal of Experimental Psychology, 65,* 368–376. doi:10.1037/h0043625

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

Cummings, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge Taylor and Francis Group.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511761676

Greene, R. L. (1989). Immediate serial recall of mixed-modality lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 266–274. doi:10.1037/0278-7393.15.2.266

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior, 10,* 562–567. doi:10.1016/S0022-5371(71)80029-4

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning & Verbal Behavior, 20,* 497–514. doi:10.1016/S0022-5371(81)90138-9

Izawa, C. (1976). Vocalized and silent tests in paired-associate learning. *The American Journal of Psychology, 89,* 681–693. doi:10.2307/1421466

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38,* 1009–1017. doi:10.3758/MC.38.8.1009

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486. doi:10.1037/a0017341

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1250–1257. doi:10.1037/a0023436

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

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

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67,* 17–29. doi:10.1016/j.jml.2012.02.004

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227–239. doi:10.1016/j.jml.2009.11.010

Kemp, F. D., & Holland, J. G. (1966). Blackout ratio and overt responses in programed instruction: Resolution of disparate results. *Journal of Educational Psychology, 57,* 109–114. doi:10.1037/h0023070

Kohl, R. M., & Roenker, D. L. (1983). Mechanism involvement during skill imagery. *Journal of Motor Behavior, 15,* 179–190.

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., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414. doi:10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. doi:10.1080/09541440701326154

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206. doi:10.3758/BF03194052

Michael, D. N., & Maccoby, N. (1953). Factors influencing verbal learning from films under varying conditions of audience participation. *Journal of Experimental Psychology, 46,* 411–418. doi:10.1037/h0063042

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

Orlando, V. P., & Hayward, K. G. (1978). A comparison of the effectiveness of three study techniques for college students. In P. D. Peterson & J. Hansen (Eds.), *Reading: Disciplined inquiry in process and practice* (pp. 242–245). Clemson, SC: National Reading Conference.

Putnam, A. L., & Roediger, M. A. (2013). The effects of response modality on retrieval. *Memory & Cognition, 41,* 36–48. doi:10.3758/s13421-012-0245-x

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 14, pp. 207–262). New York, NY: Academic Press. doi:10.1016/S0079-7421(08)60162-0

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. doi:10.1037/0033-295X.88.2.93

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

Roediger, H. L., & 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., & 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., 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* (pp. 1–36). Oxford, England: Elsevier. doi:10.1016/B978-0-12-387691-1.00001-6

Shanks, D. R., & Cameron, A. (2000). The effect of mental practice on performance in a sequential reaction time task. *Journal of Motor Behavior, 32,* 305–313. doi:10.1080/00222890009601381

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

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior, 5,* 381–391. doi:10.1016/S0022-5371(66)80048-8

Van Overschelde, J., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An expanded and updated version of the Battig and Montague (1969). norms. *Journal of Memory and Language, 50,* 289–335. doi:10.1016/j.jml.2003.10.003

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

Wohldmann, E. L., Healy, A. F., & Bourne, L. E. (2008). A mental practice superiority effect: Less retroactive interference and more transfer than physical practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 823–833. doi:10.1037/0278-7393.34.4.823

Received January 14, 2013
Revision received April 24, 2013
Accepted April 24, 2013 ■

Correction to Lohnas and Kahana (2013)

In the article “Parametric Effects of Word Frequency in Memory for Mixed Frequency Lists” by Lynn J. Lohnas and Michael J. Kahana (*Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. July 8, 2013. doi:10.1037/a0033669) there were omissions in Figure 1. All versions of this article have been corrected.

DOI: 10.1037/a0034164