Memory & Cognition
2010, 38 (8), 1009-1017
doi:10.3758/MC.38.8.1009

Enhancing visuospatial learning: The benefit of retrieval practice

Sean H. K. Kang
University of California, San Diego, La Jolla, California

Studies examining the beneficial effect of testing on memory have relied almost exclusively on verbal materials. Whether testing can improve the learning of novel, abstract visuospatial information was investigated, using Chinese characters as study stimuli. Subjects with no prior Chinese language experience studied English words paired with their Chinese equivalents. Subsequently, they either restudied the pairs twice or attempted to retrieve covertly the Chinese characters twice (with feedback provided afterward). The durations of the study and the retrieval/feedback trials were equated. On a final test given after 10 min (Experiment 1) or 24 h (Experiment 2), the subjects who had practiced retrieval were more accurate at writing/drawing the Chinese characters than were those who had studied repeatedly. The same result was replicated when learning condition was manipulated within subjects (Experiment 3). In predictions of future performance made after training, however, the subjects seemed unaware that retrieval practice was more effective than repeated studying. Testing enhances visuospatial learning, with potential implications for learning a foreign language that uses a writing script different from one's language: Repeated retrieval from memory trumps repeated studying.

Research a century old has shown that a memory test does not just measure memory but can also be a potent learning event, increasing retention of the tested information more than does additional studying (Abbott, 1909). Numerous studies since then have demonstrated this beneficial effect of testing (see Roediger & Karpicke, 2006a, for a review), and renewed interest in the testing effect during the past few years, due to growing appreciation of the potential for cognitive psychology to inform educational practice, has certainly added to the number. This effect of testing (also referred to as *retrieval practice*) has been observed for a wide range of study stimuli, including word lists (e.g., Wheeler, Ewers, & Buonanno, 2003), word definitions (e.g., Cull, 2000), prose passages (e.g., Kang, McDermott, & Roediger, 2007), foreign language vocabulary (e.g., Karpicke & Roediger, 2008), general knowledge facts (e.g., McDaniel & Fisher, 1991), and video lectures (Butler & Roediger, 2007).

One prominent theoretical account of the testing effect is that the act of retrieving information from memory strengthens the memory trace and/or makes it more accessible in the future (Bjork, 1975), with the corollary that the more effortful the retrieval at test, the greater the memorial benefit. There is converging evidence to support this retrieval effort account (e.g., Carpenter & DeLosh, 2006; Kang et al., 2007; Pyc & Rawson, 2009), and more mechanistic explications of this account have implicated elaborative retrieval; that is, effortful retrieval promotes the activation of more elaborative information, relative to less effortful retrieval or rereading, hence establishing more retrieval routes and increasing later retention. For instance, subjects are likely to activate more elaborative information when trying to retrieve the target during cued recall testing (e.g., *basket→eggs→flour→bread*) when learning weakly associated cue–target pairs (e.g., *basket–bread*) than when learning strongly associated pairs (e.g., *toast–bread*). Indeed, Carpenter (2009) found that although initial recall was poorer (and slower) during retrieval practice for weakly associated (cf. strongly associated) word pairs, final free recall of target words was better for the weakly associated pairs, providing support for this elaborative retrieval explanation (see also Chan, McDermott, & Roediger, 2006).

Since the elaborative retrieval account was proposed in the context of studies involving verbal learning, it does not seem readily applicable to the learning of nonverbal materials. From a theoretical standpoint, therefore, it is reasonable to hypothesize that the beneficial effects of testing may depend on the nature of the study stimuli. A review of the testing effect literature reveals that the studies done have relied almost exclusively on verbal materials. A handful of studies have used pictorial stimuli—Carpenter and DeLosh (2005) used faces paired with names, Glover (1989) asked students to learn the parts of a flower, and Wheeler and Roediger (1992) used pictures of common objects—but the criterial test has always required verbal responses (e.g., recall the names of the objects). The only exception is a study by Carpenter and Pashler (2007) that used two different maps containing symbols that represented man-made and geographical features. For

S. H. K. Kang, seankang@ucsd.edu

one map, subjects studied it for a period of time, followed by a period of testing, in which one item from the map would be omitted in turn and the subjects were instructed to mentally visualize the missing feature in its proper location. For the other map, the subjects studied it for a longer period of time, such that the time spent on each map in the study/test and pure study conditions was equated. After a short delay, the subjects were asked to draw both maps from memory. The finding was that the drawings were more accurate for the map that underwent testing than for the map that was only studied.

Although the results of Carpenter and Pashler (2007) indicate that the memorial benefit of retrieval practice generalizes to map learning, it is important to replicate their findings with other types of visuospatial materials. The location of features on a map can be verbally recoded (e.g., the golf course is to the west of the lake), and so it is possible that some form of verbal elaboration mediated the testing effect observed by Carpenter and Pashler. If the observed advantage of retrieval practice for map learning were due to more elaborative verbal recoding of the to-be-remembered visuospatial information, one would expect that material that is less amenable to verbal recoding would benefit less (if at all) from retrieval practice, and this would represent a critical boundary condition of the testing effect. In the literature on the memorial effects of mental rehearsal, there is a debate as to whether strategic nonverbal rehearsal of pictures is even possible (e.g., Shaffer & Shiffrin, 1972; Watkins, 1985). Recent findings by Hourihan, Ozubko, and MacLeod (2009) suggested that individuals can engage in selective rehearsal of abstract symbols. To the extent that such nonverbal mental reinstantiation and maintenance of information can occur, one might predict that retrieval practice would enhance visuospatial learning, regardless of whether the information can be verbally recoded.

The primary aim of the present study was to examine whether the testing effect would generalize to visuospatial information that is relatively difficult to verbalize. In line with this aim, subjects were presented with English words paired with their Chinese equivalents and later were tested on their recall of the Chinese characters. Unlike English, which is an alphabetic language, Chinese is logographic, with each morpheme represented by a character. Chinese characters generally resemble a square-shaped configuration of lines (referred to as *strokes*) and vary in visual complexity. To someone inexperienced with reading Chinese, the characters would appear like abstract line drawings. Examining the effect of testing on these abstract forms would therefore represent a stronger test of whether retrieval practice affects visuospatial learning. Another advantage of using Chinese characters is that they are educationally relevant stimuli, with potential practical implications for acquiring literacy in a foreign language, especially one that uses a writing system different from one's own native language.

In addition to determining the impact of testing on memory for abstract visuospatial information, a second goal was to examine the concomitant metamemorial processes that accompany testing. Specifically, when subjects monitor their learning, do they become more confident after testing (than after restudying)? In other words, do they have metacognitive awareness of the testing effect? This issue is not trivial, because the accurate monitoring of one's learning is crucial for effective self-regulation of learning (e.g., deciding which study strategy to use, which items to focus on, etc.; Dunlosky, Hertzog, Kennedy, & Thiede, 2005). Of the extant studies looking at the testing effect, only a couple have also assessed subjects' predictions of future memory performance, and both have shown (with prose materials) that subjects were more confident in predicting future memorability after rereading than after testing (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Roediger & Karpicke, 2006b). Of interest was whether this underconfidence after testing, relative to restudying, would extend to visuospatial learning.

In the present study, subjects learned Chinese characters by studying English words paired with their Chinese equivalents. For those in the restudy group, the English–Chinese word pairs were presented a total of three times for study. For those in the retrieval practice group, the pairs were presented once for study, followed by two cycles of testing with feedback, using a procedure adapted from Carrier and Pashler (1992) that equated overall processing time between the study and test conditions. During the learning phase, the subjects made predictions of their future recall of the Chinese characters. After a 10-min filled delay, the subjects were cued with the English words and were asked to write/draw the appropriate Chinese characters. To examine longer term retention, a second experiment was carried out in which the retention interval was increased to 1 day. A third experiment manipulated the learning conditions within subjects, in order to assess whether subjects' metacognitive judgments about the relative efficacy of retrieval practice vis-à-vis restudying would change when they got to experience both learning conditions.

EXPERIMENT 1

Method

Subjects. Sixty-six undergraduates from the Washington University Psychology Subject Pool participated in partial fulfillment of course requirements. All were native speakers of English and had no prior experience with Chinese languages.

Materials. Twenty Chinese characters, ranging from two to four strokes each, were selected as study stimuli. Each character was paired with its English translation (see the Appendix).

Design. Learning condition (two levels: restudy or retrieval practice) was manipulated between subjects (33 subjects in each condition). The dependent variables were the memory predictions made during the learning phase and recall performance at the final test.

Procedure. The subjects were seated at computer terminals and were informed that they would be presented with Chinese characters paired with their English translations. They were instructed to study the word pairs in anticipation of a cued recall test, in which the English word would be provided as a cue and they would be required to write out the Chinese equivalent. For all the subjects, the 20 English–Chinese pairs were presented once initially for study. A study trial consisted of a Chinese character presented in the center of the screen and the English translation directly below it. Each study trial lasted for 10 sec, with a 2-sec intertrial interval (blank screen). After an initial cycle of study, the subjects assigned to the

restudy condition were given an additional two cycles of study. The subjects in the retrieval practice condition, on the other hand, were given two cycles of testing. A test trial involved an English word cue being presented alone for 5 sec, with the instruction that the subjects should attempt to mentally visualize the appropriate Chinese character during that time, followed by the intact English–Chinese pair being presented for a further 5 sec. In this way, the restudy and retrieval practice conditions were equated in terms of total time spent during the learning phase, although the restudy group received three study cycles, whereas the retrieval practice group received one cycle of studying and two of testing. The ordering of the pairs in each study and test cycle was random. At the end of each study and test cycle, the subjects were asked to predict how many of the 20 Chinese characters they would be able to recall (i.e., write) in a cued recall test administered after 10 min (i.e., a global prediction or aggregate judgment of learning).

After the learning phase, the subjects played a video game (*Tetris*) for 10 min. They were then given a self-paced final cued recall test. Each English word appeared on the screen as a cue, and after attempting to write the equivalent Chinese character on a sheet of paper, the subjects pressed the space bar to proceed to the next word. After completing the test, the subjects were debriefed and thanked for their participation.

Results and Discussion

Global memory predictions. Due to a computer error, the predictions made by 1 subject were lost. As can be seen in Figure 1, the subjects in both the restudy and the retrieval practice groups started out with roughly equivalent predictions, and the magnitude of their predictions increased with each subsequent study/test cycle. Interestingly, the increases in predictions were greater after each subsequent study than after a test cycle. These observations were confirmed by a 2 (learning condition) × 3 (study/test cycle) mixed ANOVA (the α value for all the analyses was set at .05). Mean predictions by the restudy group were, on the whole, marginally higher than those by the retrieval practice group [$F(1,63) = 3.76$, $MS_e = .26$, $p = .057$, $\eta_p^2 = .06$]. Also, mean predictions increased as a function of study/test cycle [$F(2,126) = 50.47$, $MS_e = .54$, $\eta_p^2 = .45$]. Crucially, learning condition interacted with study/test cycle [$F(2,126) = 3.29$, $MS_e = .04$, $\eta_p^2 = .05$]. To further explore the interaction, post hoc comparisons between the restudy and retrieval practice groups were performed, using independent samples t tests. Although the mean predictions by both groups were statistically not different from each other at the first study cycle [$t(63) < 1$], the restudy group gave consistently higher predictions than did the retrieval practice group at the second and third study/test cycles [$t(63) = 2.45$, $d = 0.61$, and $t(63) = 2.14$, $d = 0.53$, respectively].

Recall performance. A written response was judged as correct (1 point) if all the strokes were present in the appropriate configuration. If a response contained an extra stroke, was missing a stroke, or had one stroke in an inappropriate position (but otherwise everything else was intact), it was judged as partially correct (1/2 point). Responses that had more than one minor error or were left blank were judged as incorrect (0). Scoring was done by a single rater. As a reliability check, responses from a random 14 subjects (21% of the total sample) were submitted to a second rater for scoring, and Cohen's kappa was .90. The left panel of Figure 2 shows the mean recall performance as a function of learning condition. The retrieval practice group exhibited reliably better performance on the final recall test than did the restudy group [$t(64) = 2.51$, $d = 0.62$].

In summary, the results demonstrated that retrieval practice, even in the absence of overt responding, enhanced learning of visuospatial information more than did equivalent restudy opportunities. These findings replicate Carpenter and Pashler (2007), who showed that the beneficial effect of testing on retention generalizes to the learning of maps, and extends the findings in the previous literature by confirming that the testing effect applies even to abstract visuospatial material that is hard to verbalize.

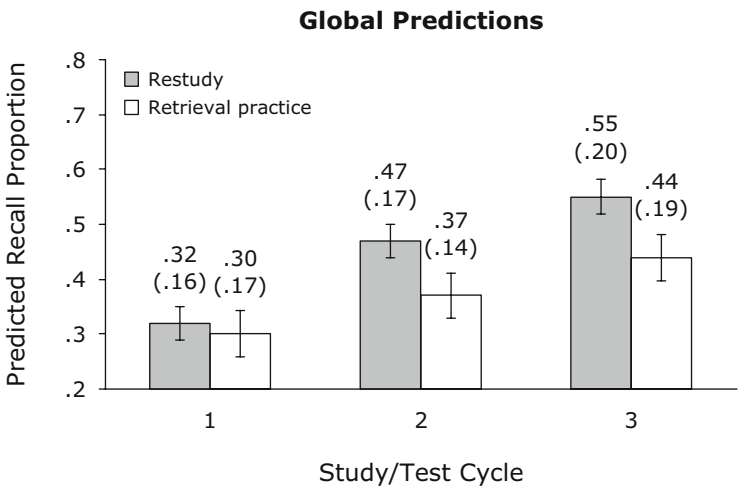

Figure 1. Mean global predictions of recall performance as a function of study/test cycle and learning condition in Experiment 1. Error bars are 95% confidence intervals. Means and standard deviations for each condition are listed above the respective bars.

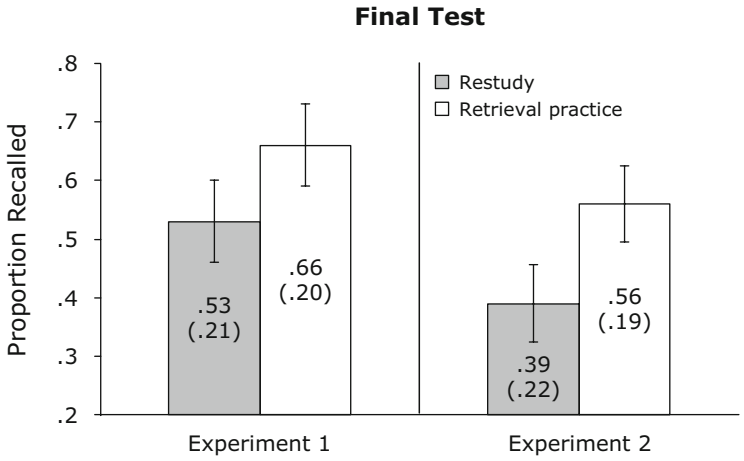

Figure 2. Mean final test recall performance for Experiment 1 (left) and Experiment 2 (right). Error bars are 95% confidence intervals. Means and standard deviations for each condition are listed in the respective bars.

Despite the robustness of the testing effect, subjects' predictions of their future memory performance were in the opposite direction—those who engaged in retrieval practice gave lower predictions than did those who studied the items repeatedly—indicating the subjects' lack of metacognitive awareness of the utility of testing for learning (Agarwal et al., 2008; Roediger & Karpicke, 2006b). This dissociation between predicted and actual performance is apparent when one visually compares Figures 1 (predictions made at the third study/test cycle) and 2 (left panel). When comparing predicted with actual performance, one might be tempted to conclude that the subjects in the restudy group were better calibrated than those in the retrieval practice group, because the global predictions by the former (at the third study cycle) were much closer to their actual recall performance than were those by the latter. However, the absolute correspondence between global memory predictions and actual test performance is not very meaningful, given the manner in which global predictions are made (see Connor, Dunlosky, & Hertzog, 1997, for a more detailed discussion). In essence, subjects are not fully sure what the final test entails, and so, when they make their global predictions, they are likely to anchor their predictions at a particular point of the scale (usually near the midpoint) and then adjust their predictions from there. The fact that the mean global predictions made by one group happened to be very close to their actual performance is inconsequential, because the absolute correspondence between predicted and actual performance would change if a different scoring criterion were used for the final test (e.g., imagine that the responses were scored in a much stricter way; recall performance would have decreased across the board and would have resulted in good correspondence between predicted and actual performance for the retrieval practice group but not for the restudy group if recall performance had decreased by ~20 percentage points).

An alternative way to determine the accuracy of subjects' monitoring of their learning is to correlate global predictions and recall performance across individuals within each group. Such a measure would index the agreement of relative ordering of individuals for the two variables (i.e., predictions and recall) and would be statistically independent of mean levels of the variables. Pearson's correlations between the global predictions at the third study/test cycle and final recall performance were .002 for the restudy group and .281 ($p = .11$) for the retrieval practice group. These correlations are low, especially for the group that engaged in retrieval practice, since prior research has shown that the correlations tend to increase with additional recall trials (e.g., Hertzog, Dixon, & Hultsch, 1990). Perhaps the fact that the to-be-remembered information was relatively abstract and hard to verbalize made it more difficult to accurately monitor its learning. In any case, the numerically higher correlation observed for the retrieval practice group (cf. restudy group) suggests that attempting retrieval of the target information may improve (albeit modestly, in this case) subjects' metacognitive monitoring.

What we can conclude from the global predictions made by the subjects is that both groups started out with their anchors at about the same point of the scale (~30%) and that the predictions changed after subsequent study/test cycles. Importantly, the adjustments in the magnitude of the predictions differed between the groups. The subjects who underwent repeated studying revised their predictions upward more than did those who underwent retrieval practice. This ordinal difference (i.e., the restudy group ended up giving higher predictions than the retrieval practice group) is interpretable and, when juxtaposed with final test performance (which exhibited the reverse ordering), reveals the disconnect between predicted and actual performance. The implications of these metamemory findings will be discussed later.

EXPERIMENT 2

The previous experiment showed that practicing retrieval of abstract visuospatial material led to better retention than did comparable study opportunities, at least with a short delay of 10 min. Although a preponderance of evidence has indicated that the testing effect is especially robust at longer retention intervals (e.g., Agarwal et al., 2008; Butler & Roediger, 2007; Roediger & Karpicke, 2006b), those studies used verbal materials, and one cannot assume that the same pattern would necessarily hold for abstract visuospatial information (Carpenter & Pashler, 2007, had a delay of 30 min). In addition, a study that examined the effect of response mode (i.e., overt vs. covert responding) in programmed instruction found that the group that made overt responses displayed an advantage over a reading control group when the retention test was delayed, but the covert responding group did not show this benefit (Krumboltz & Weisman, 1962). Therefore, it was important to ascertain whether the previously found memory advantage conferred by (covert) retrieval practice would persist with a longer delay.

Method

Subjects. Seventy-eight undergraduates from the Washington University Psychology Subject Pool, all of whom were native English speakers without Chinese language background, participated in partial fulfillment of course requirements.

Materials. The same materials were used as in Experiment 1.

Design and Procedure. The design and procedure were identical to those in Experiment 1, except that the sole dependent variable of interest was final recall performance and the retention interval between the learning phase and the final cued recall test was increased to 1 day.

Results

Recall performance. A single rater scored all the responses using the same rubric as in Experiment 1. As a reliability check, responses from a random 16 subjects (21% of the total sample) were submitted to a second rater for scoring. The interrater reliability (Cohen's κ) was .89. The right panel of Figure 2 shows the mean recall performance as a function of learning condition. As in Experiment 1, the retrieval practice group outperformed the restudy group on the final recall test [$t(76) = 3.63$, $d = 0.83$]. Although the size of the effect was numerically larger in Experiment 2 than in Experiment 1 (consistent with prior studies showing an interaction between retention interval and the testing effect), an ANOVA with combined data from Experiments 1 and 2 did not yield a significant interaction between retention interval and learning condition ($F < 1$).

EXPERIMENT 3

The previous two experiments showed that subjects who engaged in retrieval practice exhibited better learning than did those who engaged in restudying. In the next experiment, all the subjects underwent both learning conditions, each on separate halves of the stimuli, to ascertain the replicability of the testing effect on visuospatial materials in a within-subjects design. Moreover, allowing each subject to experience both learning conditions affords a stronger test of the hypothesis that subjects are generally unaware of the benefit of retrieval practice (cf. Experiment 1, in which the subjects experienced only one of the conditions). It could be that after undergoing both retrieval practice and repeated studying, subjects would become cognizant of the relative efficacy of the two learning conditions, in which case this knowledge should be reflected in predictions of future recall that are higher for the retrieval practice condition than for the repeated studying condition.

To further reduce the likelihood that any observed benefit of retrieval practice would be due to greater verbal elaboration, some of the stimuli (from the previous experiments) were replaced.[1] In addition, a questionnaire was administered at the end of the experiment to assess the subjects' strategies for learning each Chinese character.

Method

Subjects. Sixty undergraduates from the Washington University Psychology Subject Pool participated in partial fulfillment of course requirements. Again, all the subjects were native speakers of English and had no prior experience with Chinese languages.

Materials. Twenty Chinese characters, ranging from two to four strokes each, were selected as study stimuli (15 of the characters were the same as those used in the previous two experiments; refer to the Appendix for the details). Each character was paired with its English translation. The stimuli were divided into two equal sets.

Design and Procedure. The procedure was similar to that in Experiment 1, with the main difference being that learning condition (retrieval practice vs. repeated study) was manipulated within subjects. Both sets of 10 pairs each were first presented for study once (order of items was randomized). One set was then presented for two cycles of repeated studying, whereas the second set was presented for two cycles of retrieval practice. Items were blocked by learning condition (i.e., items assigned to repeated studying were restudied twice before items assigned to retrieval practice were presented, or vice versa), and the order of the learning condition was counterbalanced across subjects. After the learning phase, the subjects were asked to make two global predictions of future recall—one for each learning condition. A final cued recall test was then administered after a 10-min filled delay. Finally, the subjects were given a posttest questionnaire, which asked them to describe for each Chinese character the strategy they had used (if any) to help remember its form.

Results and Discussion

Global memory predictions. The right panel of Figure 3 shows mean predicted recall as a function of learning condition. The predicted level of recall was roughly equivalent for the retrieval practice and repeated study conditions ($t < 1$). Although this null difference may seem at variance with the results from Experiment 1 (in which the subjects who underwent retrieval practice gave lower predictions than those who underwent repeated study), both sets of results are consistent with the notion that subjects are generally unaware of the benefit retrieval practice has on learning.

Recall performance. All responses were scored by a single rater, using the same rubric as in the previous experiments. As a reliability check, responses from a random 15 subjects (25% of the total sample) were submit-

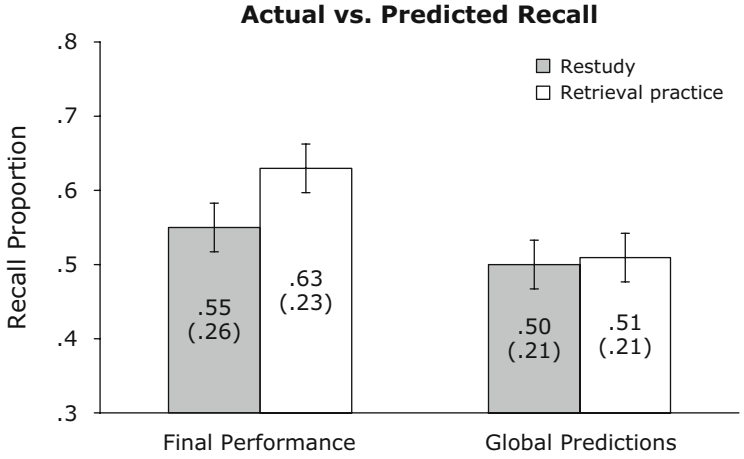

Figure 3. Mean actual and predicted final recall performance for Experiment 3. Error bars are 95% confidence intervals. Means and standard deviations for each condition are listed in the respective bars.

ted to a second rater for scoring. The interrater reliability (Cohen's κ) was .90. The left panel of Figure 3 shows the mean recall performance as a function of learning condition. As before, the retrieval practice condition yielded better recall than did the repeated study condition [$t(59) = 2.81$, $d = 0.36$].

A juxtaposition of predicted and actual recall performance suggested a dissociation between the two measures, which was supported by a marginally significant interaction between learning condition (retrieval practice vs. repeated study) and type of measure (predicted vs. actual recall) [$F(1,59) = 3.42$, $MS_e = .017$, $p = .069$, $\eta_p^2 = .055$].

Strategy questionnaire. The subjects' self-reports of their memorial strategies for each Chinese character were classified into two broad categories, depending on whether or not their descriptions mentioned any form of verbal elaboration or recoding. Included under the category of *verbal strategy* were the usage of visual analogies (e.g., the symbol for *fight* looks like two swords making sparks, the symbol for *horse* looks like a horse's face) and other forms of verbal elaboration (e.g., the symbol for *car* looks like the numeral 4, and associating that with the fact that cars have four wheels). Items for which the subjects could not verbally describe a strategy or when the listed strategy did not have a verbal component (e.g., carefully scrutinized or stared at the character) were classified under nonverbal strategy. A single rater classified the questionnaire responses, and the responses from a random 15 subjects were submitted to a second rater as a reliability check. Inter-rater agreement in the categorization of responses was very high (Cohen's $\kappa = .96$).

The majority of strategy descriptions were classified as verbal (64%), indicating that even though the stimuli were abstract visual forms, the subjects still tried to impose some sort of meaning on them by using verbal processing. Item-level mean final recall as a function of reported memorial strategy was calculated to examine whether the benefit of retrieval practice would be observed for both verbal and nonverbal strategy types. The retrieval practice condition produced better recall performance than did the repeated study condition for items learned using both verbal (.70 vs. .63) and nonverbal (.50 vs. .43) strategies. A two-way repeated measures ANOVA was performed on the item means, and it revealed main effects of strategy (verbal > nonverbal) [$F(1,19) = 38.68$, $MS_e = .019$, $\eta_p^2 = .67$] and learning condition (retrieval practice > repeated study) [$F(1,19) = 6.45$, $MS_e = .014$, $\eta_p^2 = .25$], but no interaction between the two factors ($F = 0.022$). The strategies employed by the subjects to memorize the characters were not independently manipulated, which precludes us from drawing strong inferences from these results. Nonetheless, the observed pattern in the expected direction, even in cases in which the subjects' strategies were nonverbal, suggests that the memorial benefit of retrieval practice may be independent of verbal mediation.

GENERAL DISCUSSION

The results from the three experiments demonstrate that the testing effect generalizes to abstract visuospatial stimuli. Subjects were presented with Chinese characters (paired with their English translations) to learn, and characters that underwent retrieval practice were better recalled on a subsequent test than were those that were studied an equivalent number of times. Also, the gains produced by retrieval practice were not confined merely to the short term but persisted with a longer (1-day) retention interval (Experiment 2). Due to the nature of the study stimuli (Chinese characters) and the subjects (no prior background in Chinese), the present findings suggest that verbal elaboration may not be a necessary component of the beneficial effect of retrieval practice on learning and retention and, thus, replicate and extend the findings of Carpenter and Pashler (2007).

Admittedly, the results of Experiment 3 reveal that the subjects had a preference for a verbal recoding strategy even when trying to learn abstract visual stimuli, suggest-

ing a strong tendency to use language to impose meaning when attempting to memorize abstract information. A similar trend was observed by Hourihan et al. (2009), who found that even while engaged in an articulatory suppression task, subjects reported using verbal recoding about one third of the time when encoding abstract symbols. The present findings, however, indicate that the benefit of retrieval practice does not appear to depend on the use of a verbal strategy; although it was the case that the use of a verbal strategy was associated with better final recall, the magnitude of the testing effect (at the item level) was equivalent whether a verbal or a nonverbal strategy was used during learning. These results, of course, should not be taken as evidence against the elaborative retrieval account of the testing effect. They simply suggest that there may be other mechanisms—aside from enhanced verbal elaboration or semantic activation—that can contribute to the benefit of mentally reinstating information from memory (see Wohldmann, Healy, & Bourne, 2008, for an example of mental practice aiding motor skill learning). Also, our results dovetail nicely with findings from a recent study demonstrating that elaborative encoding does not interact with retrieval practice (Karpicke & Smith, 2010), suggesting that verbal elaboration may (in some situations) be orthogonal to the factors underlying the testing effect. Given that strategy use by the subjects in the present study was not manipulated independently, it would be up to future research to disentangle the relative contributions of verbal elaboration and visual reinstatement in driving the testing effect for visual stimuli.

Another point worth noting is that the advantage of retrieval practice over restudying was observed even when no overt behavioral response was required during the learning phase. For the initial tests, subjects were instructed to mentally visualize the target Chinese character when cued with an English word (i.e., they attempted to retrieve the information covertly, without having to write it out). The beneficial effect of testing on retention thus seems to be caused by the (mental) act of retrieving information from memory, rather than the production of an outward response (e.g., Carpenter, Pashler, & Vul, 2006). Although a past study showed that covert retrieval was not as effective as overt retrieval and did not enhance long-term retention more than did a reading control group (Krumboltz & Weisman, 1962), there is also strong evidence showing that covert retrieval practice produced a significant testing effect over restudying even at delays of 2 weeks and longer (Carpenter, Pashler, Wixted, & Vul, 2008). In the present study, no overt response was elicited at the initial tests, because of the desire to equate overall processing time for the study and (initial) test trials, by using a procedure modeled after Carrier and Pashler (1992), to preclude any dismissal of the benefits of testing as being due merely to differences in exposure or processing duration. It was unfeasible to require the subjects to write out their responses while maintaining tight control over the timing of each trial. This is not to say that producing an overt response has no impact on learning. At least for Chinese reading acquisition, some have hypothesized that motor memory for the movements involved in writing Chinese characters plays an important role (Tan, Spinks, Eden, Perfetti, & Siok, 2005). Comparing covert and overt responding in order to tease apart the contributions of (mental) retrieval and writing practice is beyond the scope of the present aims and will be up to future research to resolve. The bottom line is that the present findings illustrate the utility of testing to improve retention, even when retrieval is done covertly, and suggest that testing can be implemented profitably in situations where overt responding is not convenient or possible.

The present study was not designed to examine Chinese language acquisition per se. The stimuli were selected more for their surface characteristics (i.e., abstract visuospatial forms) than from an intrinsic interest in examining the learning of Chinese characters. Nonetheless, the present findings have potential implications for foreign language acquisition. When first encountering foreign words written in a script that differs from one's native language (e.g., languages that use a writing system not based on the Roman alphabet), the words or symbols often appear as random squiggles that are relatively indistinguishable from each other. The present findings suggest that rather than repeatedly studying the foreign words, a more effective strategy for learning them would be to engage in retrieval practice by repeatedly testing oneself.

In addition to examining the effect of testing on memory, another goal of this study was to investigate how testing impacts metamemory. Global predictions of future recall performance obtained from the subjects indicated that they were not aware that retrieval practice potentiates learning. In Experiment 1, the subjects who practiced retrieving the target items gave lower predictions than did those who repeatedly studied them, in contrast to actual final test performance, which exhibited the opposite pattern of the retrieval practice group outperforming the restudy group (see also Agarwal et al., 2008; Roediger & Karpicke, 2006b). In Experiment 3, where learning condition was manipulated within subjects, comparable levels of recall predictions were obtained after retrieval practice and after repeated studying. Although the patterns of predicted recall differed in Experiments 1 and 3, both sets of results converged on the conclusion that learners are generally oblivious to the benefits of retrieval practice.

Given the present and previous metamemory findings, it is perhaps unsurprising that subjects' metacognitive judgments about testing in the laboratory are reflected also in students' study behaviors in real life, where the preferred strategy for exam preparation is rereading course notes and materials, and only a minority of students use strategies that involve retrieval practice (Karpicke, Butler, & Roediger, 2009). The challenge for educators is to get students to realize that even if the testing experience feels more effortful or difficult, it is a difficulty that has desirable consequences for long-term performance (Bjork, 1994) and, therefore, they should adopt retrieval-based study strategies in order to optimize learning and retention.

AUTHOR NOTE

Portions of this study were presented at the 50th Annual Meeting of the Psychonomic Society, Boston, MA, in November 2009. The author acknowledges Kathleen McDermott and Roddy Roediger for their helpful comments on an earlier draft of the manuscript, and Sharda Umanath, Caroline Prouvost, Robin Hong, and Kevin Chang for assisting with data scoring. This research was supported by a Collaborative Activity Award (220020041) from the James S. McDonnell Foundation. Correspondence concerning this article should be addressed to S. H. K. Kang, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109 (e-mail: seankang@ucsd.edu).

REFERENCES

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

AGARWAL, P. K., KARPICKE, J. D., KANG, S. H. K., ROEDIGER, H. L., III, & 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

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

BJORK, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

BUTLER, A. C., & ROEDIGER, H. L., III (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, **19**, 514-527. doi:10.1080/09541440701326097

CARPENTER, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **35**, 1563-1569. doi:10.1037/a0017021

CARPENTER, S. K., & DELOSH, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, **19**, 619-636. doi:10.1002/acp.1101

CARPENTER, S. K., & DELOSH, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, **34**, 268-276.

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

CARPENTER, S. K., PASHLER, H., & VUL, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, **13**, 826-830.

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.

CHAN, J. C. K., MCDERMOTT, K. B., & ROEDIGER, H. L., III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, **135**, 553-571. doi:10.1037/0096-3445.135.4.553

CONNOR, L. T., DUNLOSKY, J., & HERTZOG, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology & Aging*, **12**, 50-71. doi:10.1037/0882-7974.12.1.50

CULL, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, **14**, 215-235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

DUNLOSKY, J., HERTZOG, C., KENNEDY, M. R. T., & THIEDE, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, **10**, 4-11.

GLOVER, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, **81**, 392-399. doi:10.1037/0022-0663.81.3.392

HERTZOG, C., DIXON, R. A., & HULTSCH, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology & Aging*, **5**, 215-227. doi:10.1037/0882-7974.5.2.215

HOURIHAN, K. L., OZUBKO, J. D., & MACLEOD, C. M. (2009). Directed forgetting of visual symbols: Evidence for nonverbal selective rehearsal. *Memory & Cognition*, **37**, 1059-1068. doi:10.3758/MC.37.8.1059

KANG, S. H. K., MCDERMOTT, K. B., & ROEDIGER, H. L., III (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, **19**, 528-558. doi:10.1080/09541440601056620

KARPICKE, J. D., BUTLER, A. C., & ROEDIGER, H. L., III (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., III (2008). The critical importance of retrieval for learning. *Science*, **319**, 966-968. doi:10.1126/science.1152408

KARPICKE, J. D., & SMITH, M. (2010). *Separate mnemonic effects of retrieval practice and elaborative encoding.* Manuscript under revision.

KRUMBOLTZ, J. D., & WEISMAN, R. G. (1962). The effect of overt versus covert responding to programed instruction on immediate and delayed retention. *Journal of Educational Psychology*, **53**, 89-92. doi:10.1037/h0041100

MCDANIEL, M. A., & FISHER, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, **16**, 192-201. doi:10.1016/0361-476X(91)90037-L

PYC, M. A., & RAWSON, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory & Language*, **60**, 437-447. doi:10.1016/j.jml.2009.01.004

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

SHAFFER, W., & SHIFFRIN, R. M. (1972). Rehearsal and storage of visual information. *Journal of Experimental Psychology*, **92**, 292-296. doi:10.1037/h0032076

TAN, L. H., SPINKS, J. A., EDEN, G. F., PERFETTI, C. A., & SIOK, W. T. (2005). Reading depends on writing, in Chinese. *Proceedings of the National Academy of Sciences*, **102**, 8781-8785. doi:10.1073/pnas.0503523102

WATKINS, M. J. (1985). Strategies of picture rehearsal: A comment on Proctor's (1983) article. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 821-824. doi:10.1037/0278-7393.11.1-4.821

WHEELER, M. A., EWERS, M., & BUONANNO, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, **11**, 571-580. doi:10.1080/09658210244000414

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. doi:10.1111/j.1467-9280.1992.tb00036.x

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

NOTE

1. Reviewers of an earlier version of this article expressed the concern that some of the Chinese characters used in Experiments 1 and 2 seemed easy to recode verbally (e.g., because of some resemblance to a well-known symbol, etc.). Therefore, the stimuli were pilot tested on subjects who were asked to look at each of the 20 characters and specify whether any reminded them of other symbols that they knew or whether they could easily describe in words the form of any character. Five of the items received affirmative responses from our pilot subjects, and they were replaced for Experiment 3.

APPENDIX

fight	斗	wood	木
open	开	door	门
car	车	cow	牛
big	大	sky	天
moon	月	hand	手
up	上	book	书
fire	火	small	小
heart	心	king[a]	王
middle[a]	中	mountain[a]	山
down[a]	下	knife[a]	刀
friend[b]	友	today[b]	今
nine[b]	九	phoenix[b]	凤
horse[b]	马		

[a]Items that were used only in Experiments 1 and 2. [b]Items that were used only in Experiment 3.

(Manuscript received September 25, 2009;
revision accepted for publication May 13, 2010.)