Memory & Cognition
2006, 34 (2), 268-276

Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect

SHANA K. CARPENTER and EDWARD L. DELOSH
Colorado State University, Fort Collins, Colorado

In three experiments, we investigated the role of transfer-appropriate processing and elaborative processing in the testing effect. In Experiment 1, we examined whether the magnitude of the testing effect reflects the match between intervening and final tests by factorially manipulating the type of intervening and final tests. Retention was not enhanced for matching, relative to mismatching, intervening and final tests, contrary to the transfer-appropriate–processing view. In Experiment 2, we examined final retention as a function of the number of cues needed to retrieve items on intervening cued recall tests. In this case, fewer retrieval cues were associated with better memory on the final test. Experiment 3 replicated the findings of Experiment 2 while controlling for individual item difficulty and directly manipulating the number of cues present. These findings suggest that an intervening test may be most beneficial to final retention when it provides more potential for elaborative processing.

Memory tests are most commonly used to assess the contents of memory. Memory tests may, however, alter the very memories that they are designed to assess. One effect of tests is that they often enhance the retention of tested items, even in the absence of feedback or additional study (Allen, Mahler, & Estes, 1969; Cull, 2000; Darley & Murdock, 1971; Landauer & Eldridge, 1967; Nungester & Duchastel, 1982; Petros & Hoving, 1980; Postman & Phillips, 1961; Slamecka & Katsaiti, 1988; Wheeler & Roediger, 1992). Wheeler and Roediger found, for example, that retrieving previously presented pictures led to greater retention on a memory test given 1 week later, relative to no retrieval of the pictures. This enhancement in retention as a result of testing is often referred to as the *testing effect*.

Past research on the testing effect has shown that it cannot be accounted for by the duration of exposure to the tested material. Although recollection brings an item to mind and may, therefore, be akin to an additional presentation of the item, the testing effect is still obtained in experimental designs that directly compare memory for tested items with memory for items that are re-presented but not tested (Allen et al., 1969; Carrier & Pashler, 1992; Kuo & Hirshman, 1996, 1997). It appears, then, that there is something unique about recollection, beyond mere exposure to the material, that underlies the enhanced retention that is due to testing.

Additional research suggests that the act of retrieval may be necessary to produce an advantage of testing. Studies in which different types of intervening tests have been compared have shown that the testing effect is stronger for recall, as opposed to recognition, intervening tests (Bjork & Whitten, 1974; Glover, 1989). According to two-process models of recall (e.g., J. R. Anderson & Bower, 1972; Bahrick, 1970; Kintsch, 1970), retrieval is necessary on recall tests, but not on recognition tests. By this view, the observation that testing effects are stronger for recall than for recognition intervening tests suggests that the testing advantage may be dependent on the act of retrieval.

Two primary explanations have been offered as to why retrieval enhances memory beyond additional exposure. One explanation posits that retrieval enhances memory due to greater similarity in the processes invoked by an intervening test and final memory test, relative to an intervening study opportunity and final memory test. This transfer-appropriate–processing view (Morris, Bransford, & Franks, 1977) has received some empirical support in studies on the testing effect. McDaniel and Fisher (1991) compared participants' memory for general knowledge facts after an intervening test versus an additional study opportunity. The final memory test consisted of questions that were phrased in a way that was either similar to or different from the way in which they had been phrased on the intervening test. A comparison of the rate of final recall for items successfully retrieved on the intervening test showed that final retention was better when the test questions were phrased similarly than when they were phrased differently. This finding suggests that the testing effect may be linked to the specific retrieval cues given on the intervening test, so that tests benefit memory to the extent that the cues used on the final test match the cues given on the intervening tests.

Experiment 2 was closely based on a master's thesis submitted by the first author, portions of which were presented at the 2003 Annual Meeting of the Rocky Mountain Psychological Association, Denver, April 11–13. We thank Annie Archuletta and Nathan Castillo for their assistance in data scoring for Experiment 1. We also thank Benjamin Clegg and Alice Healy for helpful comments and conversations about this research. Correspondence concerning this article should be addressed to S. K. Carpenter, Department of Psychology, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109 (e-mail: scarpenter@psy.ucsd.edu).

A study by McDaniel, Kowitz, and Dunay (1989; see also McDaniel & Masson, 1985) used a cued-recall test consisting of either semantic or phonemic cues and arranged the conditions so that the cues provided on the intervening and the final tests were either compatible (phonemic–phonemic or semantic–semantic) or incompatible (phonemic–semantic or semantic–phonemic). McDaniel et al. (1989) found that final cued recall performance was best when the type of cue provided on the final test matched the type of cue provided on the intervening test. Final cued recall performance was not enhanced, relative to the no-test control group, when the cues given on the intervening test and the final test were not of the same type.

Although McDaniel et al.'s (1989) results support a transfer-appropriate–processing explanation, it is worth noting that there was also a trend showing that semantic cues on the intervening test produced better final retention on the final test regardless of the type of cue given on that final test. Likewise, a separate study by McDaniel and Masson (1985) reported a main effect for type of cue provided on the intervening test, so that final test performance was enhanced most by an intervening test that utilized a semantic cue. Such evidence suggests that the match between intervening and final test conditions may not provide a complete account of the testing effect.

The second explanation of the testing effect focuses on the processing that takes place at the time of the intervening test. This view proposes that because items are not readily accessible on test trials, test trials invoke more elaborative retrieval processing than do study trials (Glover, 1989; Whitten & Leonard, 1980). Supporting this elaborative-processing view, several studies have demonstrated superior retention of information under conditions that render information less accessible at the time of an intervening test. For example, a retention advantage has been observed for longer retention intervals between the presentation of items and the intervening tests (Bjork, 1988; Landauer & Eldridge, 1967; Madigan, 1969; Modigliani, 1976; Whitten & Bjork, 1977), for intervening test conditions that promote interference (Cuddy & Jacoby, 1982), and for free recall, rather than recognition, intervening tests (Bjork & Whitten, 1974; Glover, 1989).

To date, only one study has directly contrasted the transfer-appropriate–processing and elaborative-processing views in the same experiment. Glover (1989, Experiment 4) examined participants' memory for ideas from a passage by using a recognition, cued recall, or free recall test, then subsequently gave a final test that was also based on recognition, cued recall, or free recall. The types of intervening and final tests were combined factorially, so that some participants received compatible intervening and final tests, whereas others received incompatible intervening and final tests. According to Glover, intervening tests that include fewer retrieval cues (i.e., free recall and cued recall) invoke more elaborate retrieval operations than do tests that include more retrieval cues (i.e., recognition). Thus, if the testing effect reflects more elaborative retrieval processing, retention should be best for free recall intervening tests, regardless of the type of final test. If the match in processing for intervening and final tests drives the testing effect, retention should be better for compatible tests than for incompatible tests. Glover found that a free recall intervening test led to the best retention, followed by cued recall and then recognition, and that this pattern held regardless of the type of final test.

In the present study, we further examined the transfer-appropriate–processing and elaborative retrieval-processing explanations of the testing effect. Experiment 1 replicated and extended Glover's (1989) study, examining whether the testing effect reflects the compatibility of intervening and final tests, by factorially manipulating the types of intervening and final tests. In Experiment 2, we tested the elaborative-processing view by introducing a new protocol that controlled for the type of intervening test while varying the number of retrieval cues. Finally, in Experiment 3, we further tested the elaborative-processing view by controlling for individual item difficulty and directly varying the number of retrieval cues.

EXPERIMENT 1

Experiment 1 is a replication and extension of an experiment conducted by Glover (1989, Experiment 4). Glover factorially manipulated the type of intervening and final tests given (recognition, cued recall, or free recall) and found that a free recall intervening test yielded the best performance on the final test, regardless of the type of final test given. This finding provided data that were not consistent with an explanation based on transfer-appropriate processing.

In the present experiment, we also factorially manipulated the types of intervening and final tests given, using all possible combinations of recognition, cued recall, and free recall tests. Several elements of our design and procedure differed from those used by Glover (1989), however. First, Glover's no-test control condition did not include the re-presentation of nontested items; hence, the experiment did not control for exposure time across the test and no-test conditions. It is, therefore, not clear whether the testing effect observed in Glover's experiment was due to differences in exposure time or in the act of retrieval at the time of the intervening test. In our no-test control condition, nontested items received an additional study opportunity, equating exposure time across the test and no-test conditions. Second, whereas Glover manipulated the types of intervening and final tests completely between participants (requiring 12 separate groups), we used a mixed design in which the type of intervening test was manipulated within participants but the type of final test was manipulated between participants (requiring just three different groups). Third, instead of examining memory for idea units from a passage, we examined memory for eight-item word lists. Finally, we used shorter retention intervals for the intervening test (15 sec, instead of 48 h) and the final test (5 min, instead of 48 h).

As such, our experiment was an attempt to replicate Glover's (1989) study by using a paradigm that controlled for exposure time and was more amenable to laboratory study. If the testing effect is due to transfer-appropriate processing, retention should be better when there is a match, as opposed to a mismatch, between the types of intervening tests and final tests.

Method

Participants. Seventy-three undergraduate students participated in partial fulfillment of the requirements for an introductory psychology course at Colorado State University. Data from 3 participants were lost due to experimenter error. The remaining 70 participants were randomly assigned to each type of final test, so that 21 participants received a recognition final test, 23 received a cued recall final test, and 26 received a free recall final test. The participants were tested in groups of 10 or fewer.

Materials and Design. Wilson's (1988) database was used to sample a pool of 256 nouns with a frequency of occurrence greater than 10 per million and a concreteness rating between 200 and 700. Items were between three and seven letters in length and were made up of one to three syllables. One hundred twenty-eight of the items were used to construct 16 eight-item lists. Items were randomly assigned to lists, with the constraint that the items within a list were orthographically, phonologically, and semantically dissimilar and no items shared the same first letter. Four lists were designated as practice lists, 1 for each type of intervening task (recognition, cued recall, free recall, and control). The other 12 lists were experimental lists. The remaining 128 items were used as distractors for the intervening (32 items) and final (96 items) recognition tests.

Of the 12 experimental lists, 4 were tested with a recognition intervening test, 4 were tested with a cued recall intervening test, 4 were tested with a free recall intervening test, and 4 were given an additional study opportunity. For the recognition test, the participants were presented the eight items from the original list plus eight distractors and were asked to identify which of the items were old and which were new. For the cued recall test, the participants were presented with the first letter of each item, followed by blank spaces that corresponded to the number of letters in the item, and were required to retrieve the item that corresponded to the first-letter cue. To avoid confusion between items, the lists were always arranged so that no items shared the same first letter. For the free recall test, the participants were simply required to retrieve all of the items that they could remember from the original list. Finally, in the no-test control condition, all the items from a list were presented a second time for additional study. Four versions of the experiment were created, so that each list was paired with the recognition, cued recall, free recall, and control tasks and these versions were counterbalanced across groups.

All the data were collected in test booklets that contained answer sheets for the intervening tests and the final test. A separate page was included for each intervening test, with a blank filler page in between to discourage the participants from rehearsing the items from a previous test during presentation of a new word list and to discourage the participants from previewing upcoming tests. For recognition intervening tests, the 8 old and 8 new items were shown in random order, arranged in two columns. The participants were instructed to circle all of the old items. For cued recall intervening tests, the first letters of all 8 items were presented in random order, arranged in two columns, and the participants were instructed to write in the word corresponding to the first-letter cue. For free recall intervening tests, the participants were presented with a blank page and were instructed to write down as many items as possible from the list. Finally, for the no-test control condition, the 8 items were shown once again, and the participants were instructed to copy each item onto a blank page. The three types of final tests were conducted in the same fashion as the corresponding intervening tests but included the 96 test items from all 12 experimental lists.

Procedure. The participants were first instructed to read the instructions on the first page of their answer booklets. These instructions informed the participants that they would be viewing several lists of words and that they would be required to remember those words for a later memory test. The instructions also briefly described each of the four intervening task conditions, without mention of the final memory test. The experimenter then initiated the presentation of the 4 practice lists, followed by the 12 experimental lists. For each list, items were projected onto a computer screen at the front of a small classroom for 3,000 msec, with an interstimulus interval of 1,000 msec. Immediately following the presentation of the last item on a list, a distractor task was given in which eight single-digit numbers were presented sequentially at a rate of 1,000 msec per number, with a 1,000-msec interstimulus interval, for a total of about 15 sec. The participants were instructed to add the numbers together as they were being presented and then record the total in their test booklet. Following the distractor task, an on-screen cue was given (e.g., "Practice Test # 1") that corresponded to the title of the appropriate answer sheet in the participant's test booklet. When this cue was given, the participants were instructed to turn the page to the appropriate answer sheet and begin working on the test. The participants were allowed 60 sec to complete each test and were not permitted to return to the test after the time limit had elapsed. At the end of 60 sec, the participants were instructed to stop working on the test, turn the test page over, and expose only the blank filler page during presentation of the next word list. Then the next word list was presented.

After presentation of all 16 lists, the participants were given a 5-min distractor task in which they were required to write down the names of as many U.S. states as they could think of. At the end of 5 min, the participants were instructed to turn to the last page(s) of their test booklet, which corresponded to the final memory test. The recognition final test listed all of the old items in addition to 96 new distractors, arranged in random order in four columns on three separate pages, with the constraint that old items from the same list were placed at least 10 items apart. The final cued recall test listed the first-letter cues for all 96 items, arranged in random order in two columns on three separate pages, so that items from the same list were placed at least 10 items apart. The free recall final test consisted of a blank page upon which the participants were instructed to recall as many of the 96 words as they could. After reading the instructions for the final test, the participants were given 10 min to complete the test. The entire procedure lasted approximately 1 h.

Results and Discussion

Intervening test performance. For each participant, we computed the proportion of items successfully retrieved on the cued and free recall intervening tests, as well as the proportion of hits minus false alarms on the recognition intervening tests. These scores were submitted to a one-way repeated measures ANOVA, and the alpha level was set at .05. The results of the ANOVA were significant [$F(2,138) = 62.59$, $MS_e = 0.013$, $p < .05$], and post hoc comparisons using Bonferroni's correction showed that performance was better for recognition tests ($M = .89$, $SD = .11$) than for cued recall ($M = .72$, $SD = .19$) or free recall ($M = .69$, $SD = .16$) tests, with no significant difference between the latter two tests.

Final retention. We next computed the proportion of items from each intervening task condition that were retrieved on each of the three types of final tests. The means and standard errors for all the conditions are reported in

Table 1. Conducting a full factorial ANOVA on these data would involve analyses of main effects that collapse across scores on different types of tasks (e.g., control vs. recognition vs. recall). Because these tasks—particularly, tests of recognition versus recall—are believed to reflect different fundamental processes that cannot be directly compared (Kintsch, 1970; Mandler, 1980), we did not conduct a factorial ANOVA to examine main effects for the type of intervening task. Instead, we analyzed the intervening task scores separately for each type of final test, using three one-way repeated measures ANOVAs. According to the transfer-appropriate–processing explanation, we would expect final recognition performance to be higher for the recognition intervening test relative to the other intervening tasks, final cued recall performance to be higher for the cued recall intervening test relative to the other intervening tasks, and final free recall performance to be higher for the free recall intervening test relative to the other intervening tasks.

For the 21 participants who were given the recognition final test, there was no significant effect of type of intervening task on final retention [see the first row of Table 1; $F(3,60) = 0.94$, $MS_e = 0.009$, $p > .05$]. For the 23 participants who were given the cued recall final test, there was a significant effect of type of intervening task on final retention [see the second row of Table 1; $F(3,66) = 9.25$, $MS_e = 0.01$, $p < .05$]. Post hoc comparisons using Bonferroni's correction revealed that final cued recall retention was significantly higher for free recall intervening tests than for either cued recall or recognition intervening tests. For the 26 participants who were given the free recall final test, there was a significant effect of intervening task on final retention [see the third row of Table 1; $F(3,75) = 4.18$, $MS_e = 0.02$, $p < .05$]. Post hoc comparisons using Bonferroni's correction revealed that final free recall retention was significantly higher for free recall intervening tests than for recognition intervening tests.[1]

Contrary to the predictions of the transfer-appropriate–processing explanation, final test performance was not highest when the intervening and the final tests were of the same type. Instead, items tested by free recall were retained at nominally higher levels than were other intervening tasks for both the cued recall and the free recall final test conditions, and this trend was also apparent (although not significant) for the recognition final test condition. These results are similar to those observed by Glover (1989, Experiment 4), who found that free recall intervening tests yielded better final retention than did cued recall and recognition intervening tests and that this pattern held regardless of the type of final test.

Whether or not these results support the elaborative-processing view, however, remains to be determined, because recognition, cued recall and free recall tests differ in several ways not restricted to the amount of cue support (cf. Kintsch, 1970; Mandler, 1980). In Experiment 2, therefore, we examined the role of elaborative retrieval processing in the testing effect, using the same type of intervening test but varying the number of retrieval cues provided.

EXPERIMENT 2

Experiment 2 introduced a new protocol for examining the relationship between the number of cues provided on intervening tests and memory for those items on a subsequent test. On intervening tests, cues consisted of the first letter of each item, followed by blank spaces that corresponded to the number of letters in the word. The participants attempted to retrieve the item by using this first-letter cue but, if unable to recall the item, could obtain additional letters one at a time as needed. Each additional letter provided the participants with an additional cue that would be expected to boost the accessibility of the item. As such, one would expect the extent of elaborative retrieval processing to vary as a function of the number of retrieval cues given, with retrieval based on fewer cues generally requiring more elaborative retrieval processing. Thus, the present method allowed us to directly examine retention as a function of the degree of elaborative retrieval processing. If the testing effect generalizes to our new experimental protocol, final recall should be better for tested items than for studied items. Furthermore, if the magnitude of the testing effect is associated with the extent of elaborative retrieval processing, items retrieved with fewer cues should be better remembered on the final test than items retrieved with more cues.

Method

Participants. Seventy undergraduate students participated in partial fulfillment of the requirements of an introductory psychology course at Colorado State University. All the participants were tested individually on a personal computer.

Materials and Design. Wilson's (1988) database was used to sample a pool of 112 nouns with a frequency of occurrence greater

Table 1
Mean Proportion of Items Retained as a Function of Type of Intervening and Final Tests (With Standard Errors)

	Intervening Test									
	Control		Recognition		Cued Recall		Free Recall		Total	
Final Test	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Recognition	.56	.06	.53	.05	.53	.05	.57	.05	.55	.05
Cued recall	.20	.02	.14	.01	.16	.02	.29	.03	.20	.02
Free recall	.28	.04	.20	.03	.31	.05	.32	.04	.28	.04
Total	.34	.03	.28	.03	.33	.03	.39	.03		

than 20 per million and a concreteness rating between 200 and 700. Items were matched as closely as possible with regard to number of syllables (one to three) and number of letters (five to seven). In order to reduce confusion between items, given the nature of the cued recall test (see below), items were also selected so that the first two letters were always unique. The 112 nouns were randomly assigned to 14 eight-item lists, with the constraint that the items were orthographically, phonologically, and semantically dissimilar and no items within a list shared the same first letter. Two lists were practice lists that were administered at the beginning of the experiment. Of the 12 remaining lists, 6 were assigned to test trials, and 6 were assigned to study trials, and this assignment of lists to conditions was counterbalanced across participants.

Procedure. The participants first read instructions on the computer monitor. These instructions informed the participants that they would view several word lists on the computer screen and would be required to remember those words for a later memory test. The instructions briefly described both the intervening test and the study trials, without mention of the final memory test. After reading the instructions, the participants were given 2 practice lists, 1 to demonstrate the procedure for study trials and 1 to demonstrate the procedure for test trials. The 2 practice lists were followed by the 12 experimental lists. For each list, the items were presented sequentially in the center of the computer screen for 3,000 msec each, with an interstimulus interval of 1,000 msec. Immediately following the presentation of the last item in the list, the participants were given an addition distractor task for approximately 15 sec, as described in Experiment 1. Following the distractor task, the participants were given either a memory test, in the case of test trials, or an additional study opportunity, in the case of study trials. The participants did not know ahead of time whether a list would be tested or studied.

On test trials, the participants were given the first letter of each item from the most recently presented list, along with the number of blank spaces that corresponded to the number of letters in the word. If the participants were able to retrieve the word, using the first-letter cue, they were to type in the rest of the word one letter at a time. If unable to retrieve the appropriate word, the participants could obtain the second letter by pressing the space bar. The next letter appeared on the screen only when the space bar or the correct letter key was pressed on the keyboard. The participants could obtain as many letters as needed to retrieve the appropriate word, but they were encouraged to remember the word with as few letter cues as possible. Study trials were comparable to test trials in that each item from the most recently presented list was presented as a word stem consisting of the first letter and a series of blank spaces. In this case, however, the word corresponding to the stem was displayed immediately above the stem. Instead of retrieving the item from memory, the participants were simply instructed to complete the stem by typing in the letters from the displayed word.

After the 12 experimental trials had been completed, the participants were given a 5-min distractor task in which they were required to write down the names of all U.S. states that they could remember. The final memory test was then administered. The participants were instructed to turn over their answer sheets and write down as many words as they could remember from the entire experiment. Our results likely would have been contaminated by output interference had we used a cued recall test as the final test, because several word stem cues would have shared the same first letter, given that there were 96 different test items. Thus, we used a free recall test as the final test. Ten minutes were allotted to the final free recall test, and the entire experiment lasted approximately 1 h.

Results and Discussion

Intervening test performance. For each item given on intervening tests, a response was recorded each time a new letter was presented, on the basis of whether the participant entered the space bar or the correct letter key. For the word *cabin*, for example, a response was logged for each of the following stimulus presentations: *c* _ _ _ _, *c a* _ _ _, *c a b* _ _, *c a b i* _, and *c a b i n*. In this fashion, it was possible to determine the number of letters that the participants needed to retrieve each item. We assumed that an item had been retrieved once the correct letter was entered for the first time. If the participant entered the space bar for the first three presentations, followed by the correct letter on the fourth presentation (*c a b i* _), we assumed that four letters were required to retrieve the item. An examination of the data revealed that all the participants eventually entered the correct letter for all the items.

Final retention for tested versus studied items. The proportion of items correctly recalled on the final test was computed for test trials and study trials, and these scores were submitted to a one-way repeated measures ANOVA, with the alpha level set at .05. This analysis yielded a significant effect of type of trial [$F(1,69) = 120.59$, $MS_e = 0.005$, $p < .05$], so that tested items ($M = .23$, $SD = .13$) were better retained than studied items ($M = .10$, $SD = .09$).

Final retention as a function of number of cues. To examine the relationship between the number of cues provided on the intervening tests and final retention, we tabulated the number of letters needed to retrieve each of the 48 tested items for each participant. We then computed the proportion of items recalled on the final test as a function of the number of letters required. Very few participants required as many as five letters to retrieve any item, so the analysis was limited to items retrieved with one through four letters. In addition, the participants who did not contribute at least one data point to each of the four letter conditions were excluded from the analysis. Two participants did not retrieve any items with two letters, 2 additional participants did not retrieve any items with three letters, and 3 additional participants did not retrieve any items with four letters. Thus, a total of 63 participants contributed retention scores to each of the four letter cue conditions. These scores were submitted to a one-way repeated measures ANOVA, which revealed a significant effect of number of letters on final retention [$F(3,186) = 6.76$, $MS_e = 0.033$, $p < .05$]. As is shown in Table 2, retrieving items with fewer retrieval cues on the intervening test was associated with better retention on the final test.[2]

To determine whether baseline word stem production probabilities influenced the likelihood of retrieval, all of the first-letter word stems used in Experiment 2 were normed

Table 2
Mean Proportion of Items Retained as a Function of the Number of Letters Presented (With Standard Errors)

	Number of Letters							
	1		2		3		4	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Experiment 2	.29	.02	.17	.02	.17	.02	.16	.03
Experiment 3	.42	.02	.36	.03	.35	.02	.30	.02

on a separate group of 185 participants in order to obtain the probability of producing each target item without prior exposure to it. The participants in the norming study were presented with word stems one at a time for 30 sec each and were instructed to simply list all of the words that they could think of that would complete the word stem. We then compared the proportion of participants who produced each target item from the first-letter cue in the norming study with the proportion who correctly produced the item from the first-letter cue during the intervening test trials in Experiment 2. Thus, two values were associated with each of the 96 items, one representing the probability of retrieval given no prior exposure (from the norming study), and one representing the probability of retrieval given prior exposure (from Experiment 2). A correlation analysis on the 96 pairs showed that the baseline probability of producing each target item was not significantly correlated with the rate at which the item had been produced in Experiment 2 [$r = .18$; $t(94) = 1.77$, $p > .05$].

The results of Experiment 2 showed that the number of retrieval cues needed to support retrieval on intervening tests was inversely related to final retention, so that fewer retrieval cues were associated with better recall of those items on the final test. The results of Experiment 2 suggest that the benefit of tests was greatest when fewer letter cues were provided on the intervening tests. However, the method used in this experiment was such that the participants determined the number of letters provided on the intervening tests. Thus, it is not clear whether reduced cue support enhanced later retention or whether individual items requiring fewer letter cues were easier to retrieve on both the intervening and the final tests. Experiment 3 was conducted to more directly examine this issue.

EXPERIMENT 3

The purpose of Experiment 3 was to investigate the effects of number of letter cues on retention while better controlling for possible item selection artifacts. This was accomplished by using a subset of items from Experiment 2 that were of similar difficulty and by experimentally manipulating the number of letter cues provided on the intervening test trials.

Method

Participants. Sixty-three new participants were sampled from the same participant pool as that used in the previous experiments. The participants were tested individually on personal computers.

Materials and Design. We selected a subset of 32 items from Experiment 2 that had been successfully retrieved with only a one-letter cue by 65% or more of the participants. These items were arranged into four new lists of 8 items each. All the items in Experiment 3 were tested using cued recall intervening tests. We directly manipulated whether items were tested with one-, two-, three-, or four-letter cues on the intervening test trials. Two of the items in each list were cued with one letter, 2 were cued with two letters, 2 with three letters, and 2 with four letters. Four counterbalancing conditions were included so that items were tested equally often with one- through four-letter cues.

Procedure. The participants first read instructions on the computer monitor. These instructions were very similar to those in Experiment 2, except that no description of the study trial was included, because all the items in Experiment 3 were tested items. After reading the instructions, the participants were given one practice list consisting of items not included in the experimental lists. Following the practice lists, the four experimental lists were presented. For each list, items were presented sequentially in the center of the computer screen for 3,000 msec each, with an interstimulus interval of 1,000 msec. Immediately following the presentation of the last item in the list, the participants were given the addition distractor task, as described in Experiment 2. Following the distractor task, the participants were shown the word stem cues one at a time in the center of the screen, with one, two, three, or four letters shown, depending on the condition. The participants were instructed to type in the item that corresponded to each cue, followed by the Enter key, or just the Enter key alone if they were unable to remember the item. The word that the participants typed in appeared directly below the presented cue. No time limit was imposed for the intervening test trials, and no feedback was provided to verify the accuracy of retrieval. Following the last list, the participants completed the same distractor task as that used in Experiments 1 and 2, in which they were required to remember U.S. states for 5 min, and then completed a 10-min final free recall test over all of the experimental items.

Results and Discussion

We computed the proportion of items recalled on the final test as a function of the number of letters presented as cues on the intervening tests. These scores were submitted to a one-way repeated measures ANOVA, which revealed a significant effect of number of letters on final retention [$F(3,186) = 5.94$, $MS_e = 0.023$, $p < .05$]. As is shown in Table 2, retrieving items with fewer retrieval cues on the intervening test was associated with better retention on the final test.

The results of Experiment 3 fully replicate those of Experiment 2 in demonstrating that items retrieved with reduced cue support (i.e., fewer letters as cues) were better retained than were items retrieved with more cue support (i.e., more letters as cues). Unlike Experiment 2, however, Experiment 3 better controlled for the difficulty of individual items by using a subset of items that were retrieved at a high rate in the previous experiment and directly manipulated the number of letter cues provided.

GENERAL DISCUSSION

The present set of experiments yielded several key results concerning the role of transfer-appropriate processing and elaborative retrieval processing in the testing effect. Experiment 1 showed that performance was not best under conditions of matching, as opposed to mismatching, intervening and final tests. This finding replicated that obtained by Glover (1989, Experiment 4) with different memory material, shorter retention intervals for intervening and final tests, and a no-test control condition that controlled for exposure time. These results add to a growing body of evidence that the testing effect cannot be fully accounted for by the match in processes elicited by intervening and final tests. These results provide much-needed data on an explanation that has received mixed support in the literature. Due to the paucity of research investigating the testing effect from a transfer-appropriate–processing perspective, along with the past studies that do support

such an explanation (e.g., McDaniel & Fisher, 1991; McDaniel et al., 1989), further research is clearly needed in order to determine whether, and under what conditions, the transfer-appropriate–processing explanation provides the most appropriate account of the testing effect.

The elaborative retrieval hypothesis was supported by the results of Experiment 2, in which the type of intervening test was held constant and the number of letter cues varied across items. Assuming that elaborative retrieval processing generally increases as cue support decreases, we would expect that those items retrieved with fewer letter cues would be retained better than those retrieved with more letter cues. The results of Experiment 2 were consistent with this prediction, showing that the greatest proportion of items retained on the final test were those that were retrieved with a one-letter cue and that the proportion of items retained on the final test decreased as additional letter cues were added. Particularly compelling are the results of Experiment 3, in which the same pattern of results was obtained, but under conditions that better controlled for individual item difficulty and directly manipulated the number of letter cues provided for each item. Taken together, the results of all three experiments suggest that the beneficial effects of tests on memory are not always driven by the match in retrieval conditions between intervening and final tests but seem to be greatest when the intervening test conditions provide more potential for elaborative retrieval processing.

Consistent with the elaborative-processing view, a number of past studies have demonstrated that conditions designed to decrease the accessibility of an item during an intervening test often have the effect of increasing retention of that item on a later test. Better retention has been observed for free recall, as opposed to recognition, intervening tests (Bjork & Whitten, 1974; Glover, 1989); for interfering, as opposed to noninterfering, conditions during the time of an intervening test (Cuddy & Jacoby, 1982); and for longer, as opposed to shorter, retention intervals between presentation and intervening test (Landauer & Eldridge, 1967; Madigan, 1969; Modigliani, 1976; Whitten & Bjork, 1977). Thus, past and present findings converge to support the idea that a test opportunity is most beneficial when it provides the most potential for elaborative retrieval processing of items.

The elaborative retrieval hypothesis provides a somewhat more specific explanation of the testing effect, relative to earlier accounts based on trace strength—for example, the possibility that testing may produce stronger neural activity that leads to a more consolidated and durable memory trace (Cooper & Monk, 1976; Landauer & Eldridge, 1967; Whitten & Bjork, 1977) or may result in the strengthening of memory cues for long-term retention (Gotz & Jacoby, 1974; Modigliani, 1976) or the strengthening of retrieval routes to access information from long-term memory (Bjork, 1975). However, exactly what elaborative retrieval processing involves has not been clearly specified.

There are at least two possibilities that might explain the specific nature of elaborative retrieval processing in the testing effect. One possibility is that test opportunities are more likely than study opportunities to increase the variable processing of items (McDaniel & Masson, 1985). An item that is processed by two different methods (presentation followed by test), as opposed to two similar methods (presentation followed by study), may be more likely to be retrieved on a later test because that item has a greater number of cues associated with it. In support of this view, McDaniel and Masson found that test trials were more beneficial to memory retention when a different type of cue was given during presentation and the final test (phonemic–semantic and semantic–phonemic conditions), rather than when the same type of cue was given during presentation and the final test (phonemic–phonemic and semantic–semantic conditions). Presumably, the intervening test increased the variability of the target item, making it more likely to be retrieved in the context of a different cue, rather than the same cue, at the time of the final test. Along similar lines, Bjork (1975) proposed that the act of testing may create new retrieval routes, making it more likely that tested items, as opposed to studied items, will be remembered at a later time, due to a greater number of potentially effective retrieval cues.

Another possibility is that tests are more likely than study opportunities to increase the potential for item-specific processing. It has long been known that distinctive material—which is distinguished by its unique, item-specific features—is better remembered than nondistinctive material (e.g., Hunt & Lamb, 2001; Kelley & Nairne, 2001; McDaniel, DeLosh, & Merritt, 2000; Schmidt, 1991). When material is retrieved via a memory test, the act of retrieval serves to distinguish particular items from among those in the prior encoding episode, which results in the processing of the unique, item-specific features of those items (see, e.g., Kuo & Hirshman, 1997). There is also evidence that the convergence of relational and item-specific processing, which has been proposed to account for the generation effect and other memory phenomena (Hunt & McDaniel, 1993), might also apply to the testing effect. Matthews, Smith, Hunt, and Pivetta (1999) proposed that the act of retrieval relies on relational information to organize the memory search and on item-specific information to identify specific target items within that search. Although Matthews et al. did not apply this reasoning directly to the testing effect, the obvious implication is that tested items may be remembered better than studied items because the convergence of relational and item-specific information is greater in the former than in the latter.

Explanations for the testing effect that are based on variable processing and/or item-specific processing seem reasonable, given the fact that such mechanisms have been proposed to account for memory phenomena that are similar to the testing effect, such as the spacing effect and the generation effect. Studies in which the combined effects of testing and spacing have been investigated have shown that tests do not benefit memory beyond additional study if they occur at repeated intervals that are massed, rather than spaced (Carpenter & DeLosh, 2005; Cull, 2000). This ten-

dency for the testing and spacing effects to co-occur might suggest that similar mechanisms account for both of these phenomena. In several studies, the spacing effect has been investigated from a variable-processing perspective (e.g., Challis, 1993; Greene, 1989; Hintzman, 1976; Kahana & Greene, 1993), so it seems reasonable that such a perspective might also play a role in the testing effect.

The variable-processing perspective has also been applied to the generation effect in the form of the multiple-cue hypothesis proposed by Soraci and colleagues (Soraci et al., 1999; Soraci et al., 1994). According to this view, information in memory is activated during the generation of a target word, and this information acts as later retrieval cues for the target word. This extra information, even if it consists of incorrect attempts at generating the correct target, becomes associated with the correct target and provides multiple cues from which to retrieve it later on. It is possible that the processes involved in episodic retrieval that account for the testing effect are similar to those involved in semantic retrieval that account for the generation effect and that one of those processes involves the production of extra information during retrieval that later acts as multiple cues from which to retrieve target information. We have recently been investigating this possibility, and the data from several experiments in our lab suggest that multiple cues seem to play a role in the testing effect (Carpenter, 2004).

Finally, theoretical insights into the testing effect might be gained by examining the literature on retrieval-induced forgetting. M. C. Anderson and colleagues (e.g., M. C. Anderson & Neely, 1996; M. C. Anderson & Spellman, 1995) have shown that retrieval of some items can have an inhibitory effect on related, nonretrieved items, so that the future accessibility of these nonretrieved items is reduced, relative to a condition in which no retrieval took place. An interesting theoretical question for testing effect research is whether or not tested items enjoy a relative advantage over studied items because the act of retrieval enhances memory for tested items or because the act of retrieval inhibits memory for studied items. Exploring such a possibility would help determine the conditions under which tests are, and possibly are not, beneficial to memory.

Whether or not the specific pattern of results in the present study can be explained by these mechanisms remains to be addressed. Future work on the testing effect could certainly benefit from the exploration of known hypotheses that have been able to account for the pattern of results in memory phenomena that are similar to the testing effect, such as the spacing and generation effects described above. Given the amount of support for the elaborative retrieval hypothesis in the testing effect, further research into the precise mechanisms of such elaborative retrieval, as well as possible inhibitory effects of retrieval, seems warranted.

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior*, **8**, 463-470.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, **79**, 97-123.

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237-313). San Diego: Academic Press.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, **102**, 68-100.

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, **77**, 215-222.

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. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 397-401). New York: Academic Press.

Bjork, R. A., & Whitten, W. B. (1974). Recency sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, **6**, 173-189.

Carpenter, S. K. (2004). *A multiple-cue hypothesis for the testing effect*. Unpublished doctoral dissertation, Colorado State University, Fort Collins.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, **19**, 619-636.

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

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 389-396.

Cooper, A. J. R., & Monk, A. (1976). Learning for recall and learning for recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 131-156). London: Wiley.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning & Verbal Behavior*, **21**, 451-467.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, **14**, 215-235.

Darley, D. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, **93**, 66-73.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, **81**, 392-399.

Gotz, A., & Jacoby, L. L. (1974). Encoding and retrieval processes in long-term retention. *Journal of Experimental Psychology*, **102**, 291-297.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 371-377.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 10, pp. 47-91). New York: Academic Press.

Hunt, R. R., & Lamb, C. A. (2001). What causes the isolation effect? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1359-1366.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory & Language*, **32**, 421-445.

Kahana, M. J., & Greene, R. L. (1993). Effects of spacing on memory for homogeneous lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 159-162.

Kelley, M. R., & Nairne, J. S. (2001). von Restorff revisited: Isolation, generation, and memory for order. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 54-66.

Kintsch, W. (1970). Models for free-recall and recognition. In D. A. Norman (Ed.), *Models of human memory* (pp. 333-373). New York: Academic Press.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, **109**, 451-464.

Kuo, T., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory & Language*, **36**, 188-201.

Landauer, T. K., & Eldridge, L. (1967). Effect of tests without feedback and presentation–test interval in paired-associate learning. *Journal of Experimental Psychology*, **75**, 290-298.
Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning & Verbal Behavior*, **8**, 828-835.
Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, **87**, 252-271.
Matthews, T. D., Smith, R. E., Hunt, R. R., & Pivetta, C. E. (1999). Role of distinctive processing during retrieval. *Psychological Reports*, **84**, 904-916.
McDaniel, M. A., DeLosh, E. L., & Merritt, P. S. (2000). Order information and retrieval distinctiveness: The recall of common versus bizarre material. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1045-1056.
McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, **16**, 192-201.
McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition*, **17**, 423-434.
McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 371-385.
Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 609-622.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 519-533.
Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, **74**, 18-22.
Petros, T., & Hoving, K. (1980). The effects of review on young children's memory for prose. *Journal of Experimental Child Psychology*, **30**, 33-43.
Postman, L., & Phillips, L. W. (1961). Studies in incidental learning: A comparison of the methods of successive and single recalls. *Journal of Experimental Psychology*, **61**, 236-241.
Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, **19**, 523-542.
Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 716-727.
Soraci, S. A., Jr., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory & Language*, **41**, 541-559.
Soraci, S. A., Jr., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 67-78.
Wheeler, M. A., & Roediger, H. L., III (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, **3**, 240-245.
Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 465-478.
Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 127-134.
Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, **20**, 6-10.

NOTES

1. Although there are concerns about collapsing across different types of tests, we nonetheless explored potential interaction effects that might reflect an advantage for matching intervening and final test conditions by analyzing the data from Experiment 1 with a 4 × 3 (intervening task × final test) factorial mixed ANOVA, with type of intervening task as a within-participants factor and type of final test as a between-participants factor. A significant main effect of type of intervening task was observed [$F(3,201) = 9.17$, $MS_e = 0.014$, $p < .05$]. A significant main effect of type of final test was also observed [$F(2,67) = 26.65$, $MS_e = 0.107$, $p < .05$]. Importantly, the intervening task × final test interaction was not significant [$F(6,201) = 2.10$, $MS_e = 0.014$, $p > .05$], indicating that the proportion of items retained for each of the four types of intervening tasks did not vary as a function of the type of final test.

2. A similar version of Experiment 2 was also conducted in which the exposure time was similar across study and test trials by inserting a delay that prevented a new response from being typed in for at least 3 sec after the preceding response. The design of this experiment was identical in all other respects to that in Experiment 2 and yielded the same pattern of significant effects, in that final test retention was highest for items retrieved with a one-letter cue ($M = .35$, $SD = .16$), followed by two-letter cues ($M = .20$, $SD = .17$), three-letter cues ($M = .20$, $SD = .13$), and finally, four-letter cues ($M = .13$, $SD = .20$).

(Manuscript received November 11, 2003;
revision accepted for publication March 10, 2005.)