Memory & Cognition
2010, 38 (8), 1101-1109
doi:10.3758/MC.38.8.1101

Investigating the encoding–retrieval match in recognition memory: Effects of experimental design, specificity, and retention interval

Stephen A. Dewhurst and Lauren M. Knott
Lancaster University, Lancaster, England

Five experiments investigated the encoding–retrieval match in recognition memory by manipulating read and generate conditions at study and at test. Experiments 1A and 1B confirmed previous findings that reinstating encoding operations at test enhances recognition accuracy in a within-groups design but reduces recognition accuracy in a between-groups design. Experiment 2A showed that generating from anagrams at study and at test enhanced recognition accuracy even when study and test items were generated from different anagrams. Experiment 2B showed that switching from one generation task at study (e.g., anagram solution) to a different generation task at test (e.g., fragment completion) eliminated this recognition advantage. Experiment 3 showed that the recognition advantage found in Experiment 1A is reliably present up to 1 week after study. The findings are consistent with theories of memory that emphasize the importance of the match between encoding and retrieval operations.

The view that memory accuracy is determined by the match between encoding and retrieval processes is one of the cornerstones of modern memory research. Theories such as the transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977), the encoding specificity principle (Tulving & Thompson, 1973), and the procedural approach to memory (Kolers, 1973, 1975) all emphasize the importance of the encoding–retrieval match (for reviews, see Nairne, 2002; Roediger, Gallo, & Geraci, 2002; Roediger & Guynn, 1996). More recently, Kent and Lamberts (2008) suggested that the encoding–retrieval match relies on the mental simulation of encoding processes (see also Barsalou, 2008). According to these theories, records of the cognitive operations engaged at encoding are integrated with the information acquired via those operations. Reinstating the same operations at test cues the retrieval of the acquired information, leading to an increase in recognition accuracy. The aim of the present study was to identify some of the boundary conditions of this effect. Specifically, we investigated the effect of experimental design (within-groups versus between-groups manipulations), the specificity with which encoding and retrieval processes overlap, and the duration of the recognition advantage conferred by reinstating encoding operations at test.

The importance of the encoding–retrieval match has been demonstrated in a number of studies in which the orienting task carried out at study was performed again at test. For example, Glisky and Rabinowitz (1985, Experiment 1) presented participants with five-letter words that were either read intact or generated from four-letter fragments. They found that the generation effect (greater recognition accuracy for words that were generated rather than read; Slamecka & Graf, 1978) was enhanced when participants also had to generate the test items prior to the recognition decision. This finding was replicated by Dewhurst and Brandt (2007), who found that reinstatement of the generation task at test selectively enhanced the conscious recollection of studied items, as indicated by an increase in *remember* responses, but not one in *know* responses (Gardiner, 1988; Tulving, 1985).

A further demonstration of the encoding–retrieval match was reported by Engelkamp, Zimmer, Mohr, and Sellen (1994) using the enactment paradigm. Participants read a series of action phrases, such as "close the book," and on some trials were also instructed to perform the action. Consistent with the enactment effect, participants remembered enacted phrases better than they remembered phrases that were only read. Engelkamp et al. found that this effect was enhanced when participants performed the actions again prior to the recognition decision (see also Mulligan & Hornstein, 2003, who found an advantage when both participant-performed and experimenter-performed tasks were reinstated at test). Considered together, the findings above provide powerful support for the view that recognition accuracy is significantly enhanced when the operations engaged at study are reinstated at test.

A notable exception to this pattern was reported by Mulligan and Lozito (2006). They compared the effects of generation at study versus generation at test. The require-

S. A. Dewhurst, s.dewhurst@hull.ac.uk

ment to generate test items prior to a recognition decision gives rise to a phenomenon referred to as the *revelation effect*, whereby test items are more likely to be endorsed as old if they have to be revealed in order to be processed (Watkins & Peynircioğlu, 1990; see Hicks & Marsh, 1998, for a review). Mulligan and Lozito (2006) presented participants with a series of eight-letter words either intact or as anagrams, with stimulus type manipulated between groups. Test items were also presented either intact or as anagrams, again in a between-groups manipulation. Mulligan and Lozito (2006) replicated the generation effect, whereby items generated from anagrams at study were associated with greater accuracy than were items studied intact; however, test items presented as anagrams were associated with lower recognition accuracy than were test items presented intact. In other words, memory accuracy was increased by generation at study but was reduced by generation at test. As noted by Mulligan and Lozito (2006), such a pattern runs counter to the predictions of transfer-appropriate processing.

Dewhurst and Brandt (2007, Experiment 2) replicated the anagram-generation task used by Mulligan and Lozito (2006) but manipulated study and test formats in a within-groups design. Consistent with Mulligan and Lozito (2006), they found a generation effect when study items were generated from anagrams versus when they were read intact. In contrast to the findings of Mulligan and Lozito (2006), however, the generation effect was enhanced for test items that were also generated from anagrams. It is likely that the crucial difference between the two studies was the nature of the experimental design, since Mulligan and Lozito (2006) manipulated read and generate conditions between groups, whereas Dewhurst and Brandt manipulated them within groups. This interpretation is supported by the fact that the reduction in recognition accuracy reported by Mulligan and Lozito (2006) was driven mainly by an increase in false alarms, an effect that has been shown to be enhanced in a between-groups design (Hicks & Marsh, 1998; Verde & Rotello, 2004).

The first aim of the present study was to investigate the influence of experimental design on the encoding–retrieval match. Experiments 1A and 1B replicated the studies reported by Dewhurst and Brandt (2007) and Mulligan and Lozito (2006), respectively. The remaining experiments investigated some of the boundary conditions of the effect observed when read and generate conditions are manipulated within groups. Experiments 2A and 2B investigated how closely encoding and retrieval conditions must match in order to confer a recognition advantage. Experiment 2A investigated whether the advantage conferred by anagram generation at study and at test is maintained when the anagrams presented at test have a different solution key from that of the anagrams presented at study. Experiment 2B investigated whether the mnemonic advantage transfers from one generation task (e.g., anagram solution) to a second (e.g., word-fragment completion). Finally, Experiment 3 investigated the duration of the mnemonic advantage by administering the recognition test after 10-min, 24-h, 1-week, and 4-week retention intervals.

EXPERIMENT 1A

Experiment 1A was essentially a replication of Dewhurst and Brandt (2007, Experiment 2), in which participants studied eight-letter words presented either intact or as anagrams. At test, half the target items appeared in the same format as at study, and half appeared in the alternative format. On the basis of the findings of Dewhurst and Brandt and those of Glisky and Rabinowitz (1985), we expected to find a recognition advantage for words generated from anagrams at study (the generation effect) and a further advantage for the subset of those items that were generated from anagrams again at test.

Method

Participants. Forty students participated in Experiment 1A. All were native English speakers in the age range of 18–28 years. They were tested individually and received a payment of £3.

Design and Stimuli. Experiment 1A had a 2 (study format: generate, read) × 2 (test type: generate, read) repeated measures design. A set of 80 common eight-letter words was divided into two study lists, each containing 40 items. Stimuli were presented intact or as anagrams. Anagrams were created using the same method as Mulligan and Lozito (2006). Each anagram appeared on the screen with a number underneath each letter indicating its position in the solution. All anagrams had the same solution key of 54687321. Participants studied 20 anagrams and 20 intact items. Study lists were counterbalanced so that each list served as targets and lures for equal numbers of participants. The recognition test consisted of the 40 study items, with 20 presented in the same format as at study (10 as anagrams and 10 intact) and 20 in the alternative format. Forty lure items were also presented and were divided into 20 anagrams and 20 intact items. Test items were presented in a different random order for each participant.

Procedure. Study items were presented individually on an Apple Macintosh computer. The instruction "anagram" or "word" preceded each item. Each intact item remained on the screen for 2 sec, and participants were instructed to read the word aloud. Each anagram remained on the screen for 10 sec or until the participant produced the correct response, at which point the experimenter advanced the program manually. If the participant was unable to produce the correct response within the time period, the experimenter provided the answer. A practice trial consisting of three anagrams and three intact items preceded the study phase. A 10-min distractor task preceded the recognition phase. The recognition test consisted of the 40 studied items and 40 unstudied items. Half of the studied items were presented in their original study format, and half were presented in the alternative format. Participants were again asked to read aloud the intact item or to generate the correct response to the anagram before being prompted to make their recognition decision.

Results and Discussion

Mean hit and false alarm rates as a function of study and test formats are shown in Table 1, as is the discrimination measure d'. The main statistical analyses were conducted on the d' scores. In order to avoid proportions of 0 and 1 in the calculation of d', we used the correction recommended by Snodgrass and Corwin (1988) whereby .5 was added to hit and false-alarm rates, and the corrected scores were divided by $n + 1$. Since lures were neither read nor generated at study, the read–generate and generate–generate conditions used the common false alarm rate for lures that were generated at test, and the read–read and generate–read conditions used the common false alarm rate for lures that were read at test.

Table 1
Proportions of Hits and False Alarms (FAs) Plus *d′* As a Function of Study and Test Formats for Experiment 1A (Within Groups)

	Study Format									
	Generate				Read					
Test	Hits		*d′*		Hits		*d′*		FAs	
Format	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Generate	.93	.02	2.35	0.00	.73	.03	1.63	0.00	.18	.03
Read	.88	.02	1.90	0.00	.80	.02	1.58	0.00	.25	.03

The *d′* scores were analyzed in a 2 (study format: generate, read) × 2 (test format: generate, read) repeated measures ANOVA. Alpha was set at .05 in this and all subsequent analyses. A significant main effect of study format was observed, in which discrimination was greater for items that were generated at study than for those read at study [$F(1,39) = 38.32$ $MS_e = 0.28$, $\eta_p^2 = .50$]. There was also a significant main effect of test format, in which discrimination was greater for items that were generated at test than for those read at test [$F(1,39) = 5.84$, $MS_e = 0.43$, $\eta_p^2 = .13$]. The interaction was also significant [$F(1,39) = 11.01$, $MS_e = 0.16$, $\eta_p^2 = .22$]. Pairwise comparisons showed that items that were generated at study were associated with higher *d′* scores if they were generated at test than if they were read at test ($p < .05$). Test format had no reliable effect on items that were read at study ($p = .71$).

The findings of Experiment 1A are consistent with those of Dewhurst and Brandt (2007) and Glisky and Rabinowitz (1985) in showing that recognition accuracy was enhanced when the operations engaged at encoding were reinstated at test. As discussed above, Mulligan and Lozito (2006) found that generation produced opposite effects at study and at test, a finding in direct contrast to those of the aforementioned studies. Dewhurst and Brandt suggested that the different patterns of results reflected differences in experimental design, in that Mulligan and Lozito (2006) manipulated read versus generate conditions between groups, whereas the other studies used a fully within-groups design. The aim of Experiment 2B was to test this suggestion. The experiment used the same stimuli and procedure as in Experiment 1A, but—following Mulligan and Lozito (2006)—the read and generate conditions at study and at test were manipulated between groups.

EXPERIMENT 1B

Method

The method was the same as that of Experiment 1A, with the following modifications. A new group of 80 undergraduate students from Lancaster University took part. The experiment had a 2 (study format: generate, read) × 2 (test format: generate, read) between-factors design. Following Mulligan and Lozito (2006), stimuli consisted of 80 eight-letter words randomly divided into two study lists, each consisting of 40 items. Study lists were counterbalanced across participants so that they were seen as targets or lures at test equally often. The procedure followed that of Experiment 1A, with the exception that participants were randomly allocated to one of the four conditions created by crossing read and generate conditions at study and at test. For the recognition test, participants were presented with 40 study items and 40 lures. Participants who saw intact words at study followed by anagrams at test were provided with the anagram instructions and a practice phase before they began the recognition test.

Results and Discussion

Table 2 shows the mean hit and false alarm rates plus *d′* scores as a function of study and test formats. The *d′* scores were analyzed in a 2 (study format: generate, intact) × 2 (test format: generate, intact) between-factors ANOVA. A significant main effect of study format was observed [$F(1,76) = 16.03$, $MS_e = 0.28$, $\eta_p^2 = .17$], indicating that discrimination between targets and lures increased when study items were generated rather than read intact. The main effect of test was also significant [$F(1,76) = 13.51$, $MS_e = 0.28$, $\eta_p^2 = .15$], but, in contrast to the findings of Experiment 1A, discrimination was greater for items that were read intact at test than for those that were generated at test. The interaction between study and test formats was not significant ($F < 1$).

In contrast to the findings of Experiment 1A, recognition accuracy was not enhanced when the operations engaged at encoding were reinstated at test. Although a generation effect was observed when study items were presented as anagrams, presenting test items as anagrams led to a decrease in recognition accuracy. The findings of Experiments 1A and 1B therefore replicated those of Dewhurst and Brandt (2007) and Mulligan and Lozito (2006), respectively, and showed that the effect of the encoding–retrieval match depends on whether the orienting tasks are manipulated within groups or between groups.

Table 2
Proportions of Hits and False Alarms (FAs) Plus *d′* As a Function of Study and Test Formats for Experiment 1B (Between Groups)

	Study Format											
	Generate						Read					
Test	Hits		FAs		*d′*		Hits		FAs		*d′*	
Format	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Generate	.85	.02	.16	.02	2.07	0.12	.74	.03	.21	.02	1.52	0.14
Read	.86	.02	.11	.02	2.42	0.13	.85	.02	.17	.02	2.03	0.07

The key difference between Experiments 1A and 1B—and between the findings of Dewhurst and Brandt (2007) and those of Mulligan and Lozito (2006)—lies in the effect of reinstating the anagram-generation task at test. In the within-groups design employed in Experiment 1A, discrimination was enhanced when items generated at study were generated again (rather than read) at test. In contrast, in the between-groups design employed in Experiment 1B, discrimination was reduced when items generated at study were generated again at test. In order to investigate whether this difference was statistically reliable, a 2 (design: within groups, between groups) × 2 (format: generate–generate, generate–read) ANOVA was conducted on the data from the two generate-at-study conditions in Experiments 1A and 1B, using Erlebacher's (1977) procedure. Since this analysis requires equal numbers of participants in the within- and between-groups conditions, it was restricted to the first 20 participants from Experiment 1A. The analysis showed a significant interaction between design and format [$F(1,37) = 8.97$, $MS_e = 0.32$], confirming that the effect of reinstating the generation task at test is significantly influenced by experimental design.

The findings of Experiment 1A are consistent with previous findings that generating at both study and test leads to a recognition advantage beyond that obtained by generating at study alone. This raises the question of whether it is necessary to reinstate the precise operations through which the study item was generated, or whether it is sufficient simply to generate items at study and at test, regardless of the nature of the generation task. This issue was investigated by Glisky and Rabinowitz (1985, Experiments 2 and 3), who found some evidence that the encoding–retrieval match is influenced by the specificity of the overlap between the operations carried out at study and at test. They created two different fragments for each of a set of words, with one of each fragment pair presented at study. Test items were then generated either from the same fragment or from the alternative fragment. Glisky and Rabinowitz found a small but significant recognition advantage when items were generated from the same fragment at study and at test, relative to when they were generated from different fragments. Since there was no read-intact condition, however, it is impossible to say whether generating from a different fragment at test would have led to a recognition advantage, relative to a control condition. Glisky and Rabinowitz (Experiment 3) replicated the study with the addition of a read-intact condition. They again found that reinstating the same fragment at study and at test led to a recognition advantage, relative to a different fragment; however, the same-fragment condition led to a recognition advantage, relative to the read-intact condition, but the different-fragment condition did not.

The effect of manipulating the overlap between study and test items was also investigated by Gardiner, Dawson, and Sutton (1989), but in the context of a priming task. They found significant priming effects in a fragment completion task, but only when test items were identical to study items. No significant priming effect was observed when test fragments had one letter more or one letter fewer than the corresponding study fragments. In a second experiment, they found similar hyperspecificity in priming on an anagram-solution task.

Experiments 2A and 2B investigated how specific the overlap between study and test operations must be in order to produce a memory advantage in a recognition test. In Experiment 2A, we investigated whether the recognition advantage observed in Experiment 1A required the same anagrams to be presented at study and at test, or whether the effect extended to a condition in which study and test anagrams had different solution keys. In Experiment 2B, we investigated whether the advantage occurred when study and test items were generated in different ways, either from anagrams at study and from fragments at test, or vice versa.

EXPERIMENT 2A

In Experiment 2A, participants generated study items from one of two anagrams. At test, the target items were presented in the same anagram format as at study, in the alternative anagram format, or intact. Our aims were to investigate whether the advantage conferred by reinstating the encoding task was greater when the same anagram was presented at test, relative to when the alternative anagram was presented at test, and whether the alternative-anagram condition conferred a recognition advantage, relative to the read-intact condition.

Method

The method was the same as that for Experiment 1A, with the following modifications. A new group of 40 students from Lancaster University took part. All were 18–33 years of age and spoke English as their first language. The experiment followed a 2 (study format: Anagram 1, Anagram 2) × 3 (test format: anagram-same, anagram-different, read) repeated measures design. A new set of 90 eight-letter words was divided into two study sets of 45. Two anagrams were created for each word, using two different solution keys. One set of anagrams used the previous solution key of 54687321, and the second used the new key 32187546. Counterbalancing ensured that items from the study list were generated at the encoding phase, using either the first key or the second key equally often. The test list contained 90 items, 45 old and 45 new. One third of the old words were read intact at test, one third were generated at test using the same anagram key as at study, and one third were generated using the alternative anagram key. Distractor items were also divided in such a way that one third were presented as words, one third as anagrams using the first key, and one third as anagrams using the second key. The procedure was the same as that in Experiment 1A.

Results and Discussion

Table 3 shows mean hit and false alarm rates as a function of study and test formats, plus d'. Levels of correct

Table 3
Proportions of Hits and False Alarms (FAs) Plus d' As a Function of Study and Test Formats for Experiment 2A

	Hits		FAs		d'	
Test Format	*M*	*SE*	*M*	*SE*	*M*	*SE*
Anagram-same	.90	.02	.19	.02	2.19	0.12
Anagram-different	.86	.02	.20	.02	2.04	0.12
Read	.78	.02	.16	.02	1.80	0.11

or false recognition did not differ between the two anagram formats; therefore, the d' data were analyzed in a one-way (test format: anagram-same, anagram-different, read) repeated measures ANOVA. This showed a main effect of test format [$F(2,78) = 5.53$, $MS_e = 0.27$, $\eta_p^2 = .12$], and pairwise comparisons showed that discrimination was greater in both the anagram-same and the anagram-different conditions than in the read condition ($p < .05$). The anagram-same and anagram-different conditions did not differ significantly from each other ($p = .28$).

The findings of Experiment 2A show that reinstating the anagram-generation task at test increased hit rates even when the anagram differed from the one presented at study. Glisky and Rabinowitz (1985) found that the anagram-same condition produced a higher level of correct recognition than did the anagram-different condition. Although recognition accuracy in Experiment 2A was numerically higher in the anagram-same condition than in the anagram-different condition, the difference was not statistically significant. These findings contrast with those of Gardiner et al. (1989), who found that the priming of anagram solutions occurred only when study and test anagrams were identical; however, reinstating the exact anagram or fragment at test is likely to be more important in a data-driven test of the type employed by Gardiner et al. than in a test of recognition memory.

Experiment 2A showed that the recognition advantage conferred by generating from anagrams at both encoding and retrieval was reliably observed when target items were generated from different anagrams at study and at test. The aim of Experiment 2B was to investigate whether the recognition advantage was maintained when participants performed different generation tasks at study and at test. In Experiment 2B, participants generated half the study items from anagrams and half from fragments. Test items were presented in the same format as at study, in the alternative format, or intact.

EXPERIMENT 2B

Method

The method was the same as that in Experiment 2A, with the following modifications. A new group of 40 students from Lancaster University took part. All were native English speakers 18–30 years of age. The experiment followed a 2 (study format: anagram, fragment) × 3 (test format: generate-same, generate-different, word) repeated measures design. A new set of 90 eight-letter words was divided into two study lists, each containing 45 words. A fragment and an anagram were created for each word. Five-letter fragments were created by deleting the middle three letters and indicating their positions with an underscore (e.g., UMB _ _ _ LA). Each fragment had only one solution. Prior to the experiment, anagrams and word fragments were matched for accuracy rate and solution time. At study, all the items from the list had to be generated. Counterbalancing ensured that study items were generated from anagrams and fragments with equal frequency. The test list contained 90 items, 45 old and 45 new. A third of the targets were read intact, a third were generated using the same task as at study (e.g., anagrams at study and anagrams at test), and a third were generated from the alternative generation task (e.g., anagrams at study and fragments at test). A third of the distractors were presented as intact words, a third as anagrams, and a third as fragments.

Table 4
Proportions of Hits and False Alarms (FAs) Plus d' As a Function of Study and Test Formats for Experiment 2B

	Hits		FAs		d'	
Test Format	*M*	*SE*	*M*	*SE*	*M*	*SE*
Generate-same	.92	.02	.11	.02	2.54	0.12
Generate-different	.80	.03	.13	.03	2.07	0.13
Read	.82	.03	.08	.01	2.29	0.11

Results and Discussion

Table 4 shows mean hit and false alarm rates as a function of study and test formats, plus d'. Levels of correct or false recognition did not differ between the two generate conditions; therefore, the d' data were analyzed in separate one-way (test format: generate-same, generate-different, read) repeated measures ANOVAs. There was a significant main effect of test format [$F(2,78) = 6.83$, $MS_e = 0.36$, $\eta_p^2 = .15$]. Pairwise comparisons showed that discrimination was significantly greater in the generate-same condition than in both the generate-different condition and the read condition. The generate-different condition and the read condition did not differ significantly ($p = .12$).

The main finding from Experiment 2B was that generating an item at test is not in itself sufficient to enhance the recognition memory of items generated at study. The recognition advantage found in Experiments 1A and 2A was not observed when the generation task performed at test differed from the generation task performed at study. Considered together, the findings from Experiments 2A and 2B place constraints on the degree to which the reinstatement of encoding processes at test confers a recognition advantage. Experiment 2B showed that the recognition advantage is not produced simply by requiring both the study items and the test items to be generated. A recognition advantage occurs only when the same generation task is reinstated at test, although the reinstatement of the specific operations within that task (e.g., the letter transformations required to solve an anagram) is not necessary. The latter finding also constrains Kolers's (1973, 1975) argument that the probability of recognition increases with the degree of overlap between study and test operations. The degree of overlap between study and test operations was greater in the same-anagram condition than in the different-anagram condition, yet there was no significant difference between the two conditions in terms of recognition accuracy.

The experiments described above show that reinstating encoding operations at test enhances recognition accuracy, at least in a within-groups design. The final experiment in the present study investigated the duration of this effect. Experiment 3 followed the same design as Experiment 1A but included a manipulation of retention interval in which participants were tested after 10 min, 24 h, 1 week, or 4 weeks.

EXPERIMENT 3

Method

The method was the same as that in Experiment 1A, with the following modifications. A new group of 120 undergraduate students

from Lancaster University took part, with 30 participants tested after each of the four retention intervals. The experiment followed a mixed design in which retention interval was manipulated between groups and study and test formats were manipulated within groups.

Results

Table 5 shows mean hit and false alarm rates and d' scores as a function of study and test formats and retention interval. The analysis of d' showed a significant main effect of retention interval [$F(3,116) = 41.76$, $MS_e = 0.78$, $\eta_p^2 = .52$]. Planned comparisons showed that recognition accuracy decreased significantly with each increase in retention interval (all ps $< .05$). A significant main effect of study format showed that recognition was more accurate when items were generated than when they were read [$F(3,116) = 183.53$, $MS_e = 0.25$, $\eta_p^2 = .61$]. The main effect of test format was not significant ($F < 1$). There was, however, a significant interaction between study format and test format [$F(3,116) = 43.13$, $MS_e = 0.19$, $\eta_p^2 = .27$]. Pairwise comparisons showed that when items were generated at study, recognition was more accurate when items were generated at test than when they were read at test ($p < .05$). In comparison, when items were read intact at study, recognition was more accurate when items were read intact at test ($p < .05$). The three-way interaction was not significant ($F < 1$).

Separate 2 (study format: generate, read) × 2 (test format: generate, read) repeated measures ANOVAs for each retention interval showed that the enhanced recognition accuracy following generation at study was reliably present for up to 4 weeks after study. The interaction between study and test format was significant after 10-min, 24-h, and 1-week intervals but was only marginally significant after 4 weeks [$F(1,29) = 3.66$, $MS_e = 0.21$, $p = .07$, $\eta_p^2 = .11$]. Pairwise comparisons showed that the effect of reinstating the anagram condition at test was statistically significant after 10 min and after 1 week ($p < .05$) but not after 24 h ($p = .10$) or 4 weeks ($p = .18$). Reinstating the read condition at test led to significant recognition advantages after 10 min and after 24 h ($p < .05$) but not after 1 week ($p = .21$) or 1 month ($p = .48$).

An unanticipated finding from Experiment 3 was the dramatic increase in false alarms at the longer retention intervals. Statistical analysis showed a significant main effect of retention interval [$F(3,116) = 25.25$, $MS_e = 12.78$, $\eta_p^2 = .40$]. Pairwise comparisons showed that false alarms were significantly lower after 10 min than after all other retention intervals, and significantly higher after 4 weeks than after 10-min and 24-h retention intervals (all ps $< .05$). Studies of associative memory illusions, in which participants falsely remember nonstudied words after studying lists of semantic associates, have shown that false memories can be more persistent than accurate memories (e.g., Toglia, Neuschatz, & Goodwin, 1999, Experiment 2). The data from Experiment 3 indicate that the same is true even when study lists consist of unrelated words. An analysis of the response bias measure C showed that participants were more conservative after 10 min than after the longer intervals, which did not differ reliably from each other. This pattern is consistent with findings from previous research that d' increases and criterion placement becomes more liberal at longer retention intervals (e.g., Hirshman, 1995).

The results of Experiment 3 indicate that the recognition advantage produced by reinstating encoding operations at test persists for at least 1 week after study, with some evidence of an advantage after 4 weeks. The results obtained after the 10-min retention interval were broadly consistent with those obtained in Experiments 1A and 1B—specifically, a main effect of study format in which items generated at study were better recognized than were items read at study, and a significant effect of encoding–retrieval match in which reinstating encoding operations at test led to increased recognition performance. The generation effect was maintained after the 24-h, 1-week, and 4-week retention intervals, testifying to the robust nature of this effect. The study × test interaction was reliably maintained up to 1 week after study but was only marginally significant after 4 weeks. In contrast to the results of Experiment 1A, the effects of generation at test did not reach statistical significance.

Table 5
Proportions of Hits and False Alarms (FAs) Plus d' As a Function of Retention Interval and Study and Test Formats for Experiment 3

	Study Format									
	Generate				Read					
Test	Hits		d'		Hits		d'		FAs	
Format	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
10-min Interval										
Generate	.94	.02	2.30	0.12	.67	.04	1.41	0.14	.18	.02
Read	.87	.03	2.05	0.11	.78	.03	1.78	0.13	.17	.02
24-h Interval										
Generate	.93	.02	1.77	0.13	.63	.03	0.74	0.08	.35	.03
Read	.86	.03	1.54	0.13	.75	.03	1.10	0.10	.34	.03
1-Week Interval										
Generate	.88	.03	1.53	0.11	.60	.03	0.60	0.09	.37	.02
Read	.83	.03	1.13	0.10	.70	.03	0.74	0.09	.42	.03
4-Week Interval										
Generate	.77	.02	1.01	0.12	.53	.04	0.35	0.11	.40	.03
Read	.78	.03	0.81	0.12	.64	.03	0.46	0.13	.48	.03

GENERAL DISCUSSION

The findings from the present study support previous findings that recognition memory is enhanced when the cognitive operations carried out at encoding are reinstated at retrieval (Dewhurst & Brandt, 2007; Engelkamp et al., 1994; Glisky & Rabinowitz, 1985). This effect was reliably observed when read and generate conditions were manipulated in a within-groups design, but Experiment 1B confirmed the findings of Mulligan and Lozito (2006) that generation exerts opposite effects at study and at test in a between-groups design. The present findings also indicate that the encoding–retrieval match is influenced by the specificity of the overlap between study and test conditions. Generating from anagrams at both study and test increased correct recognition even when test items were

solved using anagrams different from those presented at study. In contrast, generating from different tasks at study and at test (e.g., generating from anagrams at study and from fragments at test) did not confer a recognition advantage, relative to a control condition of reading test items. Finally, Experiment 3 showed that the recognition advantage is reliably present for at least 1 week after study.

The present findings are of relevance to investigations of both the generation effect and the revelation effect. Both of these phenomena occur when participants are required to generate items, rather than simply read them. The crucial difference is that the generation effect is the result of generating items at study, whereas the revelation effect is the result of generating items at test. Although the present study investigated the effects of generation at both study and test, the generation and revelation effects have typically been studied as separate phenomena (but see Mulligan & Lozito, 2006, for a comparison of the two effects). A significant recognition advantage for words generated at study was observed in all five experiments reported here, confirming the robust nature of the generation effect in both within-groups and between-groups designs (Begg, Snider, Foley, & Goddard, 1989; see Bertsch, Pesta, Wiscott, & McDaniel, 2007, and Mulligan, 2004, for reviews). The generation effect was also reliably present up to 4 weeks after study. These findings are readily accommodated by the multifactor account of the generation effect (see, e.g., Hirshman & Bjork, 1988; Hunt & McDaniel, 1993; McDaniel, Wadill, & Einstein, 1988; Mulligan, 2001, 2004), according to which generation enhances item-specific processing of the target item, thereby increasing its distinctiveness. The multifactor account can explain generation effects in both within- and between-groups manipulations, as were found in the present study (see Mulligan & Lozito, 2004, for a review of the multifactor account of generation effects).

In contrast to the generation effect, the revelation effect was somewhat elusive in the present study. The only indication of a significant revelation effect in the analysis of d' was in Experiment 1A; however, analysis of simple main effects revealed that this effect was observed only when items had been generated at study. The revelation effect was reversed in Experiment 1B, in which study and test formats were manipulated between groups. It is possible, however, that separate effects of revelation on hit and false alarm rates may have been masked by the analysis of d'. Previous research has shown that revelation can have two effects on recognition. The basic phenomenon is a greater tendency to endorse generated items as old, which leads to an increase in hit rates. This pattern has been observed in both within- and between-groups designs (Westerman & Greene, 1996). Other studies have found that revelation leads to a reduction in memory accuracy by exerting a greater effect on false alarms than on hits, a pattern typically observed only in between-groups designs (e.g., Hicks & Marsh, 1998; Verde & Rotello, 2004). In order to investigate the presence of these effects in the present study, we conducted separate analyses of the raw hit and false alarm rates.

In the analysis of hits, the main effect of revelation at test did not reach statistical significance in Experiment 1A or at any of the retention intervals in Experiment 3. As in the analyses of d', there was a significant interaction between study and test format in which hit rates were higher when items generated at study were generated again at test. This was likely to be due to the reinstatement of the generation task, however, since generating at test did not increase hit rates for items that had been read at study. Generation at test also led to higher hit rates in Experiments 2A and 2B, but this was again likely due to the reinstatement of the generation task, since all study items were generated in one way or another. The analysis of false alarms showed a significant revelation effect only in Experiment 1B, in which study and test formats were manipulated between groups. No evidence of a revelation effect was observed when study and test formats were manipulated within groups. The false alarm data are therefore consistent with previous findings that the reduction in recognition accuracy following revelation at test is more likely to occur in a between-groups design than in a within-groups design (Hicks & Marsh, 1998; Verde & Rotello, 2004).

Why does the reinstatement of a generation task at test increase recognition accuracy in a within-groups design but decrease recognition accuracy in a between-groups design? Dewhurst and Brandt (2007) suggested that the recognition advantage in a within-groups design occurs at the level of individual items, in that reinstating the operations through which a particular item was encoded cues the recollection of that item. This is consistent with their finding that the reinstatement of encoding processes selectively enhanced correct *remember* responses. In contrast, the reversed effects of generation at test reported by Mulligan and Lozito (2006) and in Experiment 1B of the present study were driven largely by an increase in false alarms. It is likely that an encoding–retrieval match confers a recognition advantage, relative to the condition in which encoding and retrieval operations mismatch. If so, the effect is more likely to occur in a within-groups design, in which participants encounter both match and mismatch conditions. A number of influential findings in memory research have been shown to be influenced by experimental design (see McDaniel & Bugg, 2008; Mulligan & Peterson, 2008, for reviews). The present findings suggest that the same is true of the encoding–retrieval match, at least when the reinstated task involves the generation of target items.

The present findings are consistent with theories of memory that emphasize the importance of the encoding–retrieval match, such as the transfer-appropriate processing framework (Morris et al., 1977), the encoding specificity principle (Tulving & Thompson, 1973), and the procedural approach to memory (Kolers, 1973, 1975). According to such theories, reinstating encoding operations at test cues the retrieval of the information that was acquired via those operations. The findings are also consistent with the more recent view that memory retrieval is supported by the mental simulation of encoding processes (Barsalou, 2008; Kent & Lamberts, 2008). Barsalou discussed evidence that the

patterns of neural activation associated with different types of study items are reinstated when participants attempt to retrieve those items. For example, Wheeler, Peterson, and Buckner (2000) presented to-be-remembered items either visually or auditorially. When participants later retrieved those items, patterns of brain activation reflected the modalities in which the study items were presented. Kent and Lamberts also reviewed evidence from neural and behavioral studies and concluded that the encoding–retrieval match is the result of a cognitive system that simulates the processes activated during the original encoding event. The present findings can easily be accommodated within this framework by assuming that the mental simulation of the encoding event is facilitated when participants are required to perform the same orienting task at study and at test.

One discrepancy between the present findings and those of previous studies (e.g., Dewhurst & Brandt, 2007; Engelkamp et al., 1994; Glisky & Rabinowitz, 1985) is that reinstating encoding operations at test led to a recognition advantage in both the generate and read conditions. Although reinstating the read condition did not lead to a recognition advantage in Experiment 1A, this pattern was reliably present in Experiment 3 after the 10-min and 24-h retention intervals. Dewhurst and Brandt suggested that retrieval is enhanced only by the reinstatement of effortful tasks that enhance recognition memory when performed at study. They suggested that a memory advantage is less likely to occur when a relatively automatic task, such as reading, is reinstated at test. In contrast, the findings from the present study suggest that the reinstatement of a relatively automatic task can also lead to a recognition advantage. The critical factor is likely to be the discriminability of the orienting tasks, rather than their effortfulness. Support for this is provided by the results of Experiment 2A, in which recognition accuracy was enhanced only when the same task (anagram solution or fragment completion) was reinstated at test, even though both tasks led to equivalent levels of correct and false recognition when performed at encoding. This pattern is consistent with the suggestion by Nairne (2002) that it is not the encoding–retrieval match per se that determines memory performance, but rather the degree to which retrieval cues provide diagnostic information about the occurrence of a target in the study list. Repeating an encoding task at test will enhance memory to the extent that the task reinstates the distinctive features of the studied item.

To summarize, the present study investigated whether the encoding–retrieval match in recognition memory is influenced by experimental design, the specificity of the overlap between encoding and retrieval operations, and the interval between study and test. A recognition advantage was reliably observed when the orienting tasks performed at encoding and retrieval were manipulated in a within-groups design, but not when the tasks were manipulated between groups (see also Mulligan & Lozito, 2006). The recognition advantage also required the reinstatement of the particular generation task via which study items were encoded (e.g., anagram solution or fragment completion) and was eliminated when study and test items were generated in different tasks (e.g., anagrams at study and fragments at test); however, the recognition advantage did not depend on the reinstatement of the specific operations within a task (e.g., the particular letter transformations required to solve an anagram). Finally, the recognition advantage persisted up to 1 week after study but was no longer reliably present after 4 weeks. These findings delineate some of the boundary conditions of the recognition advantage conferred by the encoding–retrieval match.

AUTHOR NOTE

This research was supported by Grant RES-000-22-2294 from the Economic and Social Research Council of Great Britain. We thank the Council for its support. Correspondence concerning this article should be addressed to S. A. Dewhurst, Department of Psychology, University of Hull, Hull HU6 7RX, England (e-mail: s.dewhurst@hull.ac.uk).

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, **59**, 617-645.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 553-563.

Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 977-989.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, **35**, 201-210.

Dewhurst, S. A., & Brandt, K. R. (2007). Reinstating effortful encoding operations at test enhances episodic remembering. *Quarterly Journal of Experimental Psychology*, **60**, 543-550.

Engelkamp, J., Zimmer, H. D., Mohr, G., & Sellen, O. (1994). Memory of self-performed tasks: Self-performing during recognition. *Memory & Cognition*, **22**, 34-39.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin*, **84**, 212-219.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, **16**, 309-313.

Gardiner, J. M., Dawson, A. J., & Sutton, E. A. (1989). Specificity and generality of enhanced priming effects for self-generated study items. *American Journal of Psychology*, **102**, 295-305.

Glisky, E. L., & Rabinowitz, J. C. (1985). Enhancing the generation effect through repetition of operations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 193-205.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1105-1120.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 302-313.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 484-494.

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

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: Retrieval as mental simulation. *Trends in Cognitive Sciences*, **12**, 92-98.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition*, **1**, 347-355.

Kolers, P. A. (1975). Specificity of operations in sentence recognition. *Cognitive Psychology*, **7**, 289-306.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, **15**, 237-255.

McDaniel, M. A., Wadill, P. J., & Einstein, G. O. (1988). A contex-

tual account of the generation effect: A three-factor theory. *Journal of Memory & Language*, **27**, 521-536.

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.

MULLIGAN, N. W. (2001). Generation and hypermnesia. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 436-450.

MULLIGAN, N. W. (2004). Generation and memory for contextual details. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 838-855.

MULLIGAN, N. W., & HORNSTEIN, S. L. (2003). Memory for actions: Self-performed tasks and the reenactment effect. *Memory & Cognition*, **31**, 412-421.

MULLIGAN, N. W., & LOZITO, J. P. (2004). Self-generation and memory. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 175-214). San Diego: Elsevier.

MULLIGAN, N. W., & LOZITO, J. P. (2006). An asymmetry between memory encoding and retrieval: Revelation, generation, and transfer-appropriate processing. *Psychological Science*, **17**, 7-11.

MULLIGAN, N. W., & PETERSON, D. (2008). Assessing a retrieval account of the generation and perceptual-interference effects. *Memory & Cognition*, **36**, 1371-1382.

NAIRNE, J. S. (2002). The myth of the encoding–retrieval match. *Memory*, **10**, 389-395.

ROEDIGER, H. L., III, GALLO, D. A., & GERACI, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory*, **10**, 319-332.

ROEDIGER, H. L., III, & GUYNN, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Vol. 10. Memory* (pp. 197-236). San Diego: Academic Press.

SLAMECKA, N. J., & GRAF, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, **4**, 592-604.

SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

TOGLIA, M. P., NEUSCHATZ, J. S., & GOODWIN, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, **7**, 233-256.

TULVING, E. (1985). Memory and consciousness. *Canadian Psychology*, **26**, 1-12.

TULVING, E., & THOMPSON, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352-373.

VERDE, M. F., & ROTELLO, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review*, **11**, 560-566.

WATKINS, M. J., & PEYNIRCIOĞLU, Z. F. (1990). The revelation effect: When disguising test items induces recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 1012-1020.

WESTERMAN, D. L., & GREENE, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1147-1153.

WHEELER, M. E., PETERSON, S. E., & BUCKNER, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences*, **97**, 11125-11129.

(Manuscript received July 17, 2009;
revision accepted for publication May 8, 2010.)