Memory & Cognition
2001, 29 (6), 798-805

Demonstrations of a generation effect in context memory

ELIZABETH J. MARSH
Washington University, St. Louis, Missouri

and

GABRIEL EDELMAN and GORDON H. BOWER
Stanford University, Stanford, California

Generation often leads to increased memorability within a laboratory context (see, e.g., Slamecka & Graf, 1978). Of interest in the present study is whether the benefits of generation extend beyond item memory to context memory. To investigate this question, in three experiments, we asked subjects to remember in which of two contexts they had read or generated words. In Experiment 1, the contexts were two different rooms; in Experiment 2A, the contexts were two different computer screens; in Experiment 2B, the contexts were different perceptual characteristics of the to-be-remembered words. In all experiments, subjects were better at remembering the context of generated words than of read words.

Memories originate from many different sources. Although many of our memories are based on actual perceptions and experiences, other memories are formed through mental operations such as thinking, dreaming, and imagining. Internal generation of events often leads to increased memorability within a laboratory context. People are better able to remember words generated in response to cues than words that they have merely read (the *generation effect*; Slamecka & Graf, 1978; see also Jacoby, 1978). Similarly, people are better at judging how often they have imagined a picture or generated a word than how often they have seen it (Johnson, Raye, Wang, & Taylor, 1979; Raye, Johnson, & Taylor, 1980). In some circumstances, generation may reduce memory errors; people are less likely to misattribute a novel word to their own prior generation than to a previous presentation by an external source (the *it-had-to-be-you effect*; Johnson, Raye, Foley, & Foley, 1981).

Of interest in the present paper is whether the benefits of generation extend beyond memory for the item itself to memory for contextual information. Does generating an item increase one's ability to remember not only the generated item, but also the background spatiotemporal context in which it was generated? Or does generation reduce contextual memory, perhaps as subjects focus their attention on themselves and the to-be-generated events rather than on the events' context? As we will describe, the literature currently does not provide clear answers to these questions.

Context as Spatiotemporal Background

Most relevant to the present research are studies in which context is defined as spatiotemporal background, such as the room in which a word has been studied. Such studies have yielded mixed results on the effects of generation on context memory. In general, such studies have been done within the *source monitoring framework* (Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981). Johnson and colleagues have hypothesized that memories of perceived versus imagined events differ in prototypical ways, and that these differences form the basis of successful source monitoring. A memory of a perceived event allegedly contains more inherent sensory and spatiotemporal information than does a memory of an imagined event. In contrast, imagined events are predicted to contain more information about the mental processes involved in their creation. Source misattributions occur when a memory trace has characteristics that are atypical for its source. For example, an easily generated event may be misattributed to perception because of its vividness and lack of cognitive operations (see, e.g., Finke, Johnson, & Shyi, 1988; Foley, Durso, Wilder, & Friedman, 1991; Johnson et al., 1981; Johnson et al., 1979).

In two experiments, Johnson and colleagues found some support for their hypothesis that perceived memories

We thank Patrick Dolan and Jessica Lang Nowinski for many useful discussions and for comments on the manuscript, and Larry Jacoby and Heidi Sivers for comments on an earlier draft of the manuscript. We are grateful to Rich Marsh and two anonymous reviewers for many helpful comments and suggestions. We thank Roddy Roediger for his encouragement during the writing process. The first author is supported by an NRSA postdoctoral fellowship (1F32MH12567) from the National Institute of Mental Health. Preparation of the manuscript was facilitated by funds from a Stanford University Undergraduate Research Opportunity grant to the second author, and by Grant MH-47575 from the National Institute of Mental Health to the third author. Correspondence should be addressed to E. J. Marsh, Department of Psychology, Washington University, One Brookings Drive, Campus Box 1125, St. Louis, MO 63130-4899 (e-mail: emarsh@artsci.wustl.edu).

would contain more contextual information than would imagined memories (Johnson, Raye, Foley, & Kim, 1982). In their first experiment, subjects studied 48 items, with half presented on a screen to the subjects' right and the other half presented on a screen to the subjects' left. One half of the items were perceived (line drawing plus its name) and one half were imagined (from name only). The subjects were later tested for both old–new recognition memory (of the words) and left–right location memory, both immediately and after a 1-week delay. A conventional test of significance yielded no significant effect of item type (imagined vs. perceived) on location memory. However, a sign test suggested that subjects' immediate memories for locations were slightly better for seen pictures (85%) than for imagined ones (80%); this effect was not significant at the delay.

In their second experiment, Johnson et al. (1982) found similarly weak support for the hypothesis that generated memories would be less associated with temporal context. Subjects saw or imagined each item in a series of line drawings. They then received a temporal order test. Each test page contained eight perceived or eight imagined items (in word format) that were to be labeled with the numbers one to eight to indicate relative temporal order. Two dependent measures were considered, mean position ranking and the number of test pages on which items from each of the 8 study positions received a correct rank. Overall performance was quite low; only 19% of items were assigned to the correct position, and items were generally attributed to the middle of the studied list ($M = 4.5$ of a possible 8 positions). There was no difference in mean position ranking between imagined and perceived items, and this did not interact with study position. Similarly, there was no main effect of item type (perceived vs. imagined) on mean number of correct rankings. When these data were examined across list positions, there was a significant difference (via a sign test) between imagined and perceived events only for those that had been studied in the last one eighth of the list. Thus, out of 16 possible comparisons (8 study positions for each of 2 dependent measures), only 1 yielded a significant difference between imagined and perceived events.

Adding to the confusion is a study finding opposite results—namely, that internally generated memories were *more* associated with background context (Koriat, Ben-Zur, & Druch, 1991, Experiment 1). Koriat et al. had subjects study 40 high-frequency words, half of which were presented in an enclosed lab room and half in an office with a window. Of the 20 words shown in each room, half were fragments to be mentally completed by the subject, and half were words presented intact. The generate versus read tasks were intended to create, respectively, relatively more internal versus external memory conditions. In a third room, subjects were tested for old–new recognition memory and context memory (room where an old test word had been presented). Context (room) memory was significantly better for generated than for read words.

Context as Audience

Context is not limited to spatiotemporal details but also includes one's audience; the backdrop of item generation can include other people in addition to peripheral room characteristics. At first glance, studies in which context is defined as audience provide the strongest evidence that generation impairs context memory (e.g., Jurica & Shimamura, 1999; Koriat et al., 1991, Experiment 2). However, as we will argue in the following paragraphs, these studies are not conclusive on this point, because they confounded generated versus presented events with two different dependent measures.

To explicate this argument, let us begin with a description of what we consider to be the most complete study examining people's ability to remember audience (context) following generation versus presentation, that of Brown, Jones, and Davis (1995). Their study phase was modeled after group conversation. Subjects took turns reading categories (questioner role), generating category exemplars (responder role), or simply watching the question-and-answer process (bystander role). The subjects participated in groups of four, and each one took all three conversational roles during the course of the experiment. At test, the subjects recalled the response generated for each of the category cues identified as old, and then identified who asked the question (questioner identification) and who answered it (responder identification). For the purposes of the present paper, of primary interest is subjects' ability to identify who asked the question, because one's questioner (or audience) is part of the background context of response generation. As is shown in Table 1, the critical result was that questioner identification was the same across all conversational roles, both immediately and after a delay. Having answered the question did not make questioner identification any better or worse; generation had no effect on context memory when it was operationalized as audience. However, it is also critical to note the occurrence of a main effect of the type of source task. In all conditions, the subjects found questioner identification to be much more difficult than responder identification (see Table 1).

Other studies concluding that generation impairs context memory are hard to interpret for the following reason: They compared the equivalent of questioner identification following generation with responder identification following presentation. As the Brown et al. (1995) study shows, questioner identification is more difficult than responder identification. If only generation is paired with the more difficult source task, questioner identification, then it is difficult to interpret findings of negative generation effects.

For example, Jurica and Shimamura (1999) presented faces paired with either questions ("What type of music do you like to listen to?") or statements ("The type of music I like to listen to is jazz"). Thus, generating events involved answering questions, whereas reading events involved viewing statements. At test, subjects first recalled the topics. They then took a recognition test that re-presented the

Table 1
In Conversational Paradigms That Operationalize Context as Audience, the Responder/Actor Role Is Analogous to a Generation Condition, Whereas the Bystander Role Parallels a Read Condition

	Subject's Role in Experiment		
Study	Responder/Actor (Generation)	Bystander (Read)	Questioner
Brown et al. (1995, Exp. 1), immediate			
Questioner ID	65.8	70.6	61.8
Responder ID	99.1	94.1	96.2
Brown et al. (1995, Exp. 1), delayed			
Questioner ID	37.0	32.5	35.4
Responder ID	87.9	57.7	62.6
Jurica & Shimamura (1999, Exp. 1)			
Questioner ID	58.2	–	–
Responder ID	–	76.5	–
Koriat et al. (1991, Exp. 2)			
Audience ID	84.1	–	–
Actor ID	–	93.6	–

Note—The data from Brown et al. (1995) show that questioner/audience identification is (1) harder than responder/actor identification, but that (2) it does not vary as a function of role. The remaining studies do not contain all the cells necessary for one to conclude that generation impairs context (audience) memory. Values represent percentage correct identifications.

statements and questions. For items that were identified as old, the subjects indicated which face had been paired with each item. As was expected, a generation effect was obtained in recall. However, a negative generation effect was found for context memory. As is shown in Table 1, the subjects were not good at identifying the questioners (58.2%); however, it is not known whether this would have varied with conversational role, because those conditions were not included in the experiment. Because the subjects in this experiment only answered questions and read statements, it is unknown how asking or overhearing questions would have affected questioner identification. Thus, no conclusions can be drawn about the effects of generation on audience memory. The subjects were better at identifying who had made statements (76%) than who had asked questions (58.2%). This is not surprising, given the Brown et al. (1995) finding that responder identification is an easier task than questioner identification. Thus, the poorer performance in the generation condition could have been due either to generation or to the difficulty of the source task, questioner identification.

A similar argument could be made regarding Koriat et al. (1991, Experiment 2). In the first phase of that experiment, subjects performed and watched actions with a partner; in the second phase of the experiment, they switched partners and performed and watched another set of actions in an identical room. They were then asked to attribute old actions to either the first or the second phase of the experiment. Presumably subjects based their phase decisions on their partners, since this was the most salient difference between the two phases of the experiment (the only other difference being time). For performed actions, this was equivalent to a responder's or an actor's memory for the audience, a difficult judgment in the Brown et al. (1995) study. Judging watched actions was equivalent to probing a bystander's memory for someone else's response, an easier judgment in the Brown et al. study. The results are shown in Table 1. Although phase identification was worse for performed actions, it is unclear whether this was due to generation or to the requirement to remember audience.

The Present Research

Previous research does not provide a clear answer to the question of how generation affects memory for context. When context has been defined as spatiotemporal background (e.g., the study room), effects have been found in both directions. When context was defined as audience, generating a response did not help or impair one's memory for audience in the only study that completely crossed all conversational roles (Brown et al., 1995).

The following experiments were aimed at understanding under what circumstances subjects would show better memory for contextual information. Our subjects read and generated category exemplars in contexts similar to those used by Koriat et al. (1991) and Johnson et al. (1982). They were later asked to remember where they had studied each word. In Experiment 1, we considered the possibility that the conflicting results might have occurred because of differences in the stimuli used previously. A positive generation effect was found only by Koriat et al., who used very typical category exemplars as the to-be-remembered material. Highly typical exemplars are often easy for subjects to generate (e.g., Johnson et al., 1981), and thus, the generation task might not have reduced subjects' ability to focus on other things such as the room context. A negative generation effect may only occur when generation is difficult, creating an inward attentional focus. To test these ideas, we manipulated the taxonomic typicality of the to-be-remembered items in our first experiment.

Surprisingly, in Experiment 1 we found that internally generated memories were *more* associated with context, regardless of the difficulty of generation. These results were replicated and extended in Experiments 2A and 2B. In Experiment 2A, we examined the association of read and generated words to a different kind of context (viz., adjacent computer screens). In Experiment 2B, we examined the association of read and generated items to a different kind of perceptual information (viz., color and font of the visually presented words).

EXPERIMENT 1
Room Memory

Method

Subjects. Thirty-two Stanford University undergraduates participated in the experiment in exchange for monetary compensation.

Contexts. There were two study phases in the experiment, each conducted in a different location. The *lab room*, a 12 × 9 ft rectangular room without windows and illuminated by fluorescent lights, contained two computers, a table, three chairs, and two file cabinets. One of the computers was used to present the to-be-remembered stimuli. The *lounge* was an open space with high ceilings and large windows that revealed a view of Stanford's flowered gardens. This space was furnished with couches; the subject sat on one of the couches facing the experimenter and the windows. The experimenter presented the stimuli by flipping through a series of 8 × 11 in. pages. One half of the subjects began the experiment in the lab and then proceeded to the lounge; the remaining subjects began the experiment in the lounge and then proceeded to the lab.

Stimuli. The stimuli were 120 words, 80 of which were studied. The same 40 words always served as lures, whereas the other words were rotated through experimental conditions as was appropriate. The words were drawn from 20 categories from the Battig and Montague (1969) norms, with 6 exemplars from each category. Four words from each category were studied, and the remaining 2 words served as lures on the memory test. Of the 4 studied words, 2 were presented to the subject (with the category cue) for reading. For the other 2 words, subjects were presented with the category cue plus the first two letters of the word, and they were required to (silently) generate the word completion (Slamecka & Graf, 1978). Thus each word was presented with its category label, regardless of whether it was read or generated.

The typicality of studied exemplars was manipulated so that some words would be easier to generate than others, as was done by Johnson et al. (1981). For both studied and new items, one half were high-frequency (typical) category exemplars and one half were lower frequency (less typical) exemplars. Words were judged as high typicality if they appeared among the seven top responses in the Battig and Montague (1969) category norms, and as low typicality otherwise. All words were pretested to ensure that (1) generation was possible in the time allotted and (2) subjects overwhelmingly generated the desired exemplar.

All subjects read and generated items in both the lounge and the lab. Of the four studied category exemplars, two were studied in the lounge (1 read, 1 generated) and two were studied in the lab (1 read, 1 generated). In each context, subjects studied one typical and one less typical exemplar from each category; if the high- (or low-) typicality exemplar was read in the first context, then its mate was generated in the second context. The 20 categories were divided into two sets (A,B) so that item type (e.g., high-typicality generated word) was not confounded with context: for example, if a high-typicality exemplar from set A was generated in the lab, a high-typicality exemplar from set B was generated in the lounge. Thus, eight different study lists were created in order to counterbalance three variables: item type (read vs. generated), study context (lab vs. lounge), and pairings of words with a context. Within a study list, presentation was randomized, either via the computer program (in the lab) or by shuffling the study cards (in the lounge).

Procedure. All subjects arrived at the lab room for the experiment; one half were led to the lounge to begin the first study session. In the lab room, the Superlab program was used to present the stimuli on a Macintosh computer. In the lounge, stimuli were presented on 8 × 11 in. sheets of paper secured by binder rings; the experimenter paced the subject through the experiment by turning the pages in the binder. In both the lab and the lounge, words were presented for 5 sec each (timed by the computer in the lab, by stopwatch in the lounge). The printed words appeared in the same font and character size on the printed pages as on the computer screen. In both contexts, subjects were instructed not to read or generate aloud. Following the first study phase, the experimenter led the subject to the second learning room. The same procedure was followed in the second study phase, except with the remaining 40 study words.

Following the study phase, subjects filled out unrelated questionnaires for 30 min; none of the to-be-remembered words appeared in these filler tasks. After the delay, subjects took a paper-and-pencil memory test. The test included all 120 words, in their completed form, without the category labels. These 120 words represented the 40 generated words (20 from each room), the 40 read words (20 from each room), and 40 new distractors. The subjects identified each item as having been presented in the lab room, the lounge, or neither (new). Three different orders of the test were used. The tests were created so that no more than three old or three new responses appeared in a row, and so that generated, read, and new items were spread out evenly. The subjects were never asked to indicate whether items had been read or generated during study.

Finally, to check on subjects' ability to generate the critical words, they were re-presented with the category cues and stems of the generated words, and asked to write in the last part of the word using the same generated word that they had used during the study phase. The subjects were asked to put a line through a word if they had been unable to generate it during the study phase. Following completion of this task, the subjects were debriefed and paid. The entire experiment took less than 1 h to complete.

Results

Unless otherwise noted, results were significant at the .05 alpha level.

Recognition memory. A 2 (room: lab or lounge) × 2 (category typicality: high or low) × 2 (item type: generated or read) ANOVA model was conducted on mean proportion of old items correctly labeled as "old." As is shown in Table 2, there was a main effect of item type [$F(1,31) = 137.4, MS_e = 0.03$]. Replicating the often found generation effect, subjects were better at recognizing items they had generated as opposed to items they had read (Slamecka & Graf, 1978).

Main effects of category typicality and room were not significant. Subjects were equally accurate in recognizing words that were more or less typical of the category [$F(1,31) = 2.33, MS_e = 0.01$], and were equally accurate when recognizing words from the lounge and the lab ($F < 1$). The only significant interaction was between typicality and item type [$F(1,31) = 9.79, MS_e = 0.02$]. For generated words, subjects were equally good at recognizing more and less typical category exemplars. In contrast, subjects were better at recognizing less typical category exemplars as opposed to more typical ones.

There were few false alarms: on the average, less than two per participant. There was no significant effect of category typicality on false alarms ($F < 1$). Subjects were

Table 2
Mean Proportion of Items Correctly Recognized and Context Memory Scores (Conditional on Correct Recognition) in Experiment 1

	Old–New Recognition		Context Memory	
Context	Generated	Read	Generated	Read
Lab Context				
High category typicality	.88	.59	.74	.68
Low category typicality	.86	.69	.73	.70
	.87	.64	.74	.69
Lounge Context				
High category typicality	.91	.58	.85	.75
Low category typicality	.86	.63	.82	.73
	.88	.61	.84	.74
Mean	.88	.63	.79	.71

equally likely to attribute false alarms to the lab and the lounge ($F < 1$).

Context memory. Context identification scores were computed by dividing the total number of times that context was correctly identified by the number of items recognized as old, regardless of context. A 2 (room) × 2 (category typicality) × 2 (item type) ANOVA model was tested on these context identification scores.

Item type produced a main effect [$F(1,31) = 10.24$, $MS_e = 0.03$]. As is shown in Table 2, subjects were better at recognizing the context for old items that they had generated (79%) than for old items that they had read (71%). Although subjects were also better at identifying when items had been studied in the more distinctive lounge context [$F(1,31) = 4.15$, $MS_e = 0.09$], item type and room did not interact [$F(1,31) = 1.06$, $MS_e = 0.04$]. There was no main effect of category typicality, nor were any of the other interactions significant (Fs < 1).

Discussion

In Experiment 1, subjects read and generated category exemplars in two distinct contexts, a lounge and a lab. The subjects showed better recognition memory for generated words than for read words, a result that is consistent with the generation effect (Slamecka & Graf, 1978). The subjects were also significantly better at remembering *where* they had *generated* words than at remembering where they had *seen* words. While this result is similar to the findings in the first experiment of Koriat et al. (1991), it is contrary to the findings of Johnson et al. (1982).

The manipulation of exemplar typicality had no effect on context memory; thus, in the remaining two experiments we focused on the more interesting result that generated items were more associated with context than were read items. Because several procedural aspects differed between our studies and prior work, in Experiments 2A and 2B we sought to replicate our main finding, using two very different context manipulations. In Experiment 2A, as in Johnson et al.'s (1982) Experiment 1, subjects saw and generated items on screens to their left and right all in intermingled order, so that contexts were not temporally distinct. In Experiment 2B, we used a very different context manipulation: color. Subjects were required to remember the color and font in which they had studied items. Similar results across all three experiments would allow us to claim with confidence that generated events are more associated with contextual information.

EXPERIMENT 2A
Left–Right Memory

Method

Subjects. Sixteen Stanford University undergraduate students participated in the experiment in exchange for monetary compensation.

Contexts. Two computer monitors were set up on a table next to each other in the lab room used in Experiment 1. Half of the words were presented on the screen on the left monitor, and the other half were presented on the screen of the right monitor. The two contexts were identical except for their spatial location.

Stimuli. The stimuli were 90 words, 60 of which were studied. The words were drawn from 15 of the 20 categories used in Experiment 1. The memory load was reduced in Experiment 2, because pilot data demonstrated that less distinctive contexts had a detrimental effect on location memory. As in Experiment 1, half of the studied words were read and half were generated.

As in Experiment 1, eight different study lists were created to counterbalance item type (read vs. generated), context (left vs. right screen), and word pairing. The order of the study words was randomized for each subject.

Procedure. The entire experiment took place in the lab room. The experiment began with the single study phase. The experimenter read the instructions aloud to each subject, during which they were told to pay attention to the words in preparation for a memory test. They were not informed that they would later need to make left/right discriminations. All subjects studied a single block of 60 words, 30 of which were read and 30 of which were generated. These stimuli were presented on two identical Macintosh computers using the Superlab program. The screens were placed 3 in. apart so that subjects had to turn their heads slightly to see the stimuli on each screen. Each word was presented for 5 sec. The subjects were instructed to study silently. The first category cue and exemplar pair was presented on the left screen, the second on the right, and subsequent presentations alternated back and forth. When words appeared on one screen, the other screen remained blank. Half of the words presented on each screen were read, and the other half were generated. The order of the read and generate trials was randomized. Thus, although the subjects knew that the words alternated in a left–right pattern, they never knew whether a given trial would involve reading or generating.

After the study phase, the subjects carried out unrelated filler tasks for 2 min. Following these tasks, a paper-and-pencil memory test was administered. As in Experiment 1, the subjects responded to each of

90 words in their completed form, without category cues. Thirty words had been presented on the left screen (one half generated and one half read), 30 words had been presented on the right screen (one half generated and one half read), and 30 words had not been presented. For each word, the subjects made an old–new judgment; they labeled a word as "old" if it had been presented in the experiment. For words labeled as "old," they then indicated on which monitor (left or right) it had been presented.

Results

Recognition memory. A 2 (context) × 2 (item type) ANOVA was conducted on mean number of items correctly identified as "old." The results are shown in Table 3. Replicating Experiment 1, subjects recognized more items that they had generated (93%) than items that they had read (78%) [$F(1,15) = 12.21, MS_e = 0.01$]. Subjects were equally good at recognizing words from the two contexts, and context and item type did not interact (Fs < 1).

False alarms were not analyzed because they were very rare, averaging less than one per subject.

Context memory. As in Experiment 1, context identification scores were calculated conditional upon old recognition. A 2 (context) × 2 (item type) ANOVA model was conducted on context identification scores. Most important was a main effect of item type [$F(1,15) = 18.65$, $MS_e = 0.02$]. As is shown in Table 3, subjects were better at recognizing the context (left vs. right computer screen) for items that they had generated (87%) than for items that they had read (79%). Subjects were equally good at remembering the two contexts [$F(1,15) = 1.29, MS_e = 0.01$, $p > .2$], and context did not interact with item type ($F < 1$).

EXPERIMENT 2B
Color Memory

Method

Subjects. Sixteen Stanford University undergraduate students participated in the experiment in exchange for monetary compensation.

Context and Stimuli. Half of the words were presented on the screen in an orange, all uppercase print font. The other half of the words were presented in a blue, all lowercase script font. These redundant cues were used to make the two presentation modes more salient. All words were presented in the same central spot on the computer screen. The stimuli were the same as in Experiment 2A.

Table 3
Recognition and Context Memory Scores for Experiments 2A (Left–Right Memory) and 2B (Blue–Orange Memory)

	Old–New Recognition		Context Memory	
Experiment	Generated	Read	Generated	Read
Experiment 2A				
Left	.93	.75	.87	.81
Right	.94	.80	.88	.77
	.93	.78	.87	.79
Experiment 2B				
Blue	.86	.71	.75	.67
Orange	.89	.75	.70	.65
	.88	.73	.73	.66

Four study lists were created. The words were fully counterbalanced for color (blue vs. orange) and item type (read vs. generated). In this experiment, words were not counterbalanced for word pairings.

Procedure. The procedure was almost identical to that of Experiment 2A; the two experiments differed only in the type of to-be-remembered contextual information (location vs. color). As in Experiment 2A, the subjects were told that their memories would be tested but were not informed regarding the nature of the upcoming memory test. All 60 study words were presented in a single block, on a single Macintosh computer, using the Superlab program. Each word was presented for 5 sec; presentation order was randomized for each subject.

After the study phase, the subjects received 2 min of unrelated filler tasks. They then took a paper-and-pencil memory test. As in Experiments 1 and 2A, the subjects judged each test item in its completed form, without category cues. For each word, they decided whether it had been presented in the study list (was "old"). For items judged "old," they made an additional judgment of whether the item had been presented in orange or blue letters. The subjects circled their responses on the memory test.

After the subject completed the memory test, the experimenter circled all old items incorrectly identified as "new." The subject then went back over the form and made a color judgment for those circled items.

Results

Recognition memory. A 2 (item type) × 2 (context) ANOVA model was conducted on proportion of items correctly labeled as old. As is shown in Table 3, subjects recognized more generated (88%) than read (73%) words [$F(1,15) = 12.76, MS_e = 0.03$]. There was no main effect of context, and the interaction between context and item type was not significant. Again, false alarms were rare (averaging less than one per subject), and so these were not analyzed.

Context memory. As before, context identification scores were calculated conditional upon old recognition and a 2 (item type) × 2 (context) ANOVA model was performed on these scores. The effect of item type failed to reach significance [$F(1,15) = 2.62, MS_e = 0.02, p > .1$] but reached conventional levels of significance when a one-tailed t test was used to compare context memory for read versus generated items [$t(15) = 1.77, p < .05$]. As is shown in Table 3, subjects were better at recognizing the context for items they had generated (72%) than for items they had read (65%). There was no effect of context, and the interaction between context and item type was not significant (Fs < 1).

When subjects were forced to make context decisions about the old items they had incorrectly labeled as new, performance was near chance (46% correct).

Discussion of Experiments 2A and 2B

Despite our making the two contexts less distinctive and using a very similar method to that of Johnson et al. (1982), the subjects in Experiment 2A were still better at identifying where they had generated a word than where they had seen a word. Similarly, the subjects in Experiment 2B showed better memory for the color and font of internally generated as opposed to read words. Although this effect is by only a one-tailed t test, the effect is of the

same size and direction as the significant context effects found in the first two experiments.

GENERAL DISCUSSION

The present research was motivated by the conflicting results from prior studies examining the effects of generation on context memory. In all three of the present experiments, a positive generation effect was found: Memory for context was better for words that had been generated during the study phase. Subjects were better at remembering the context of generated words regardless of whether that context was a room (Experiment 1), a computer monitor (Experiment 2A), or a color (Experiment 2B).

Why might generation increase not only memory for the to-be-remembered item, but also memory for the spatiotemporal context in which the item has been studied? In some ways, it is not surprising that internally generated events in laboratory paradigms might be associated with more types of information than previously predicted. As we described in our introduction, results based on different behavioral measures already suggest a kind of "privileged status" for self-generated events (e.g., Raye et al., 1980; Slamecka & Graf, 1978). We can think of two ways in which the benefits of generation might extend to context memory. First, it might be that generation, a deep and elaborative encoding, leads to the binding of many features into the memory trace (see Chalfonte & Johnson, 1996, for a discussion of binding). However, it is important to note that we are not making any claims about the nature of these bound features. For example, when we say that people are better at remembering that a word was generated in response to a blue cue, we are not saying anything about how vividly they had imagined that target word. Second, the act of generation may cause stronger memories to be laid down (see Hoffman, 1997, and R. L. Marsh & Bower, 1993, for discussions of the item-strength account of source monitoring), and possibly this would include all aspects of the event occasioning the generation. At present, we are not able to separate these two possible accounts for our effects.

We will now discuss the implications of our results for (1) how we conceptualize memories for internally generated events, (2) source monitoring, and (3) the larger literature on generation and context memory. We will begin with a joint discussion of the characteristics of internally generated memories and the role that these play in source monitoring, and then we will discuss the implications of our results for the literature reviewed in our introduction.

At first glance, our results may seem to contradict those of studies in which self-reports have been used to evaluate the characteristics of internally generated events. As we have described already, Johnson and colleagues argue that internally generated memories contain less perceptual and spatiotemporal information, and that these kinds of qualitative differences between memories allow for successful source monitoring (e.g., Johnson et al., 1993; Johnson & Raye, 1981). People's self-report data support this characterization; people rate actual autobiographical memories as containing more sensory and contextual information than do memories of imagined events (Hashtroudi, Johnson, & Chrosniak, 1990; Johnson, Foley, Suengas, & Raye, 1988; Suengas & Johnson, 1988). People report basing judgments about whether or not their memories are of events that have actually occurred on these typical differences between real and imagined memories (Johnson & Suengas, 1989; but see Schooler, Gerhard, & Loftus, 1986). Yet although these results may seem to contradict our findings, we believe they do not, for two reasons. First, we begin by noting again that we are not saying anything about the vividness of the features tested in our paradigms. We are simply stating that for whatever reason (e.g., binding or item strength), generation strengthened the association between the to-be-remembered item and its context. This context may or may not be vividly encoded, but even memory for low vividness is informative, because memories may be attributed on the basis of not only the presence of characteristics but also the lack of them (e.g., see Finke et al., 1988; R. L. Marsh & Hicks, 1998). Second, we also acknowledge that our stimuli (read and generated words) are very different from the autobiographical memories rated by Johnson's subjects. It may be that the characteristics of generated events depend on the nature of the stimuli; we will return to this theme later in the discussion.

Although our simple experiments yielded easily interpretable data, we are aware that we have really made the larger picture less clear. Although our experiments all clearly showed positive generation effects in context memory, there remain findings in the opposite direction (e.g., Johnson et al., 1982). In Experiment 1, we tested one factor that we thought might affect the relationship between generation and context memory—namely, the difficulty of the generation task. Difficulty of generation turned out not to be the critical factor, at least not at the levels of difficulty used in our experiment. More research will be needed to successfully predict when a generation effect will versus will not appear in context memory. One possibility follows from the item-strength account of source monitoring (R. L. Marsh & Bower, 1993). Although internally generated events often yield stronger representations in memory than do externally presented events (R. L. Marsh & Landau, 1995), there are situations in which external memories represent the stronger class (Hoffman, 1997). If stronger memories are stronger with respect to all aspects of an event, including its spatiotemporal context, then a positive generation effect should occur in context memory when generated events form the stronger class in memory. According to this view, external memories may be more associated with contextual information in situations in which they represent the stronger class of memories, such as when the source discrimination is made between actual pictures and imagined pictures (Johnson et al., 1982) or between real and imagined autobiographical events (Johnson et al., 1988). This view predicts that any operation that enhances memory for the item should enhance memory for its context of presentation. Such ideas remain to be tested experimentally.

REFERENCES

BATTIG, W. F., & MONTAGUE, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, **80** (3, Pt. 2), 1-46.

BROWN, A. S., JONES, E. M., & DAVIS, T. L. (1995). Age differences in conversational source monitoring. *Psychology & Aging*, **10**, 111-122.

CHALFONTE, B. L., & JOHNSON, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition*, **24**, 403-416.

FINKE, R. A., JOHNSON, M. K., & SHYI, G. C. W. (1988). Memory confusions for real and imagined completions of symmetrical visual patterns. *Memory & Cognition*, **16**, 133-137.

FOLEY, M. A., DURSO, F. T., WILDER, A., & FRIEDMAN R. (1991). Developmental comparisons of explicit versus implicit imagery and reality monitoring. *Journal of Experimental Child Psychology*, **51**, 1-13.

HASHTROUDI, S., JOHNSON, M. K., & CHROSNIAK, L. D. (1990). Aging and qualitative characteristics of memories for perceived and imagined complex events. *Psychology & Aging*, **5**, 119-126.

HOFFMAN, H. G. (1997). Role of memory strength in reality monitoring decisions: Evidence from source attribution biases. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 371-383.

JACOBY, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning & Verbal Behavior*, **17**, 649-667.

JOHNSON, M. K., FOLEY, M. A., SUENGAS, A. G., & RAYE, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, **117**, 371-376.

JOHNSON, M. K., HASHTROUDI, S., & LINDSAY, D. S. (1993). Source monitoring. *Psychological Bulletin*, **114**, 3-28.

JOHNSON, M. K., & RAYE, C. L. (1981). Reality monitoring. *Psychological Review*, **88**, 67-85.

JOHNSON, M. K., RAYE, C. L., FOLEY, H. J., & FOLEY, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology*, **94**, 37-64.

JOHNSON, M. K., RAYE, C. L., FOLEY, M. A., & KIM, J. K. (1982). Pictures and images: Spatial and temporal information compared. *Bulletin of the Psychonomic Society*, **19**, 23-26.

JOHNSON, M. K., RAYE, C. L., WANG, A. Y., & TAYLOR, T. H. (1979). Fact and fantasy: The roles of accuracy and variability in confusing imaginations with perceptual experiences. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 229-240.

JOHNSON, M. K., & SUENGAS, A. G. (1989). Reality monitoring judgments of other people's memories. *Bulletin of the Psychonomic Society*, **27**, 107-110.

JURICA, P. J., & SHIMAMURA, A. P. (1999). Monitoring item and source information: Evidence for a negative generation effect in source memory. *Memory & Cognition*, **27**, 648-656.

KORIAT, A., BEN-ZUR, H., & DRUCH, A. (1991). The contextualization of input and output events in memory. *Psychological Research/Psychologische Forschung*, **53**, 260-270.

MARSH, R. L., & BOWER, G. H. (1993). Eliciting cryptomnesia: Unconscious plagiarism in a puzzle task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 673-688.

MARSH, R. L., & HICKS, J. (1998). Test formats change source-monitoring decision processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1137-1151.

MARSH, R. L., & LANDAU, J. D. (1995). Item availability in cryptomnesia: Assessing its role in two paradigms of unconscious plagiarism. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1568-1582.

RAYE, C. L., JOHNSON, M. K., & TAYLOR, T. H. (1980). Is there something special about memory for internally generated information? *Memory & Cognition*, **8**, 141-148.

SCHOOLER, J. W., GERHARD, D., & LOFTUS, E. F. (1986). Qualities of the unreal. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 171-181.

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

SUENGAS, A. G., & JOHNSON, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. *Journal of Experimental Psychology: General*, **117**, 377-389.

(Manuscript received July 1, 1999;
revision accepted for publication May 18, 2001.)