

Output monitoring tests reveal false memories of memories that never existed

Richard L. Marsh

University of Georgia, USA

Jason L. Hicks

Louisiana State University, USA

Three experiments were conducted to assess participants' beliefs about potential false memories that might have occurred during free recall tests. An input–output monitoring test was administered that required participants to discriminate between items that were studied and recalled, studied and not recalled, or were entirely new. Critical lures from Roediger and McDermott's (1995) paradigm were inserted into this test. The results demonstrated that participants believed erroneously recalled items were both studied and recalled. The intriguing finding was that *unrecalled* items were believed to have been studied approximately 80% of the time, and half of those were also believed to have been recalled. This result represents a dual false memory effect in which items were believed to have been studied and also to have been recalled. The ramifications of this new procedure are discussed in terms of proposed experiments that might clarify the genesis of these false memories.

What a person believes to be true about his or her memory is quite important. Because such beliefs determine expectations, confidence, and trust in what one recollects, they are particularly significant in the context of false memories. In general, false memories occur when one reports that an event was experienced in a particular situation despite the event never having occurred at all under those circumstances. Although there are several laboratory techniques for studying false memories (see Payne & Blackwell, 1998), none has received as much attention as Roediger and McDermott's (1995) revitalisation and extension of Deese's (1959) laboratory paradigm. That technique involves presenting lists of semantically related words such as *mad*, *fear*, *hate*, *temper*, and *rage*. In tests of free recall or recognition memory, the claim that nonpresented critical lures appeared on the studied list (e.g., *anger*) constitutes one form of an experimentally induced false memory. The various encoding and testing

conditions that give rise to false memories using this technique have been well documented in a number of recent investigations (e.g., Gallo, Roberts, & Seamon, 1997; Lampinen, Neuschatz, & Payne, 1999; McDermott, 1996; McDermott & Roediger, 1998; Read, 1996; Robinson & Roediger, 1997; Schacter, Israel, & Racine, 1999; Seamon, Luo, & Gallo, 1998; Sommers & Lewis, 1999). Only a few of these reports, however, have seriously analysed what participants believe about their false memories. Therefore, our goal in the present study is to reintroduce an older technique for examining beliefs about memories in order to assess aspects of the phenomenological experience of false remembering.

Of course, Roediger and McDermott's (1995) paradigm does provide some indication of what people believe about their false memories depending on the manner in which their memory is tested. For example, memory theorists generally assume that participants truly believe that a free

Requests for reprints should be sent to Richard L. Marsh, Department of Psychology, University of Georgia, Athens, GA 30602-3013, USA. Email: rlmars@uga.edu or jhicks@lsu.edu

We wish to thank Jeffrey Sachs and Lorie Ritschel for their dedicated help in collecting the data.

recalled item was presented on a studied list. Given that extra-list intrusions in free recall are rather rare (e.g., Bower & Mann, 1992), this assumption is quite reasonable. Consequently, when a person erroneously recalls *anger*, a researcher can be fairly confident that the participant believes *anger* appeared earlier in the experiment. Unfortunately, the same assumption cannot be made as easily when participants claim that an item such *anger* or *mad* is old on a recognition test. In the recognition context, participants may guess that items are old, thereby causing some studied items to be correctly identified without necessarily strong beliefs of having experienced them earlier. In order to assess beliefs concerning a given item, postrecognition judgements such as confidence ratings, remember-know decisions, or source attributions must be made in conjunction with each item on the recognition test.

Such judgements have demonstrated that participants label a majority of the nonpresented critical lures as distinctly "remembered" (Roediger & McDermott, 1995) as opposed to "known", and they will also willingly attribute almost all of the erroneously identified items to a particular source (Hicks & Marsh, 1999; Mather, Henkel, & Johnson, 1997; Payne, Elie, Blackwell, & Neuschatz, 1996). These results from augmented standard recognition tests suggest that participants hold strong beliefs of having experienced items that were never actually presented in that context. There are, however, other ways to assess those beliefs apart from the measures that have been used to date. For example, if one believes that an item was recollected on a previous occasion, then one's confidence that the item was truly experienced prior to that recollection may be reinforced. Therefore, if one claims that *anger* was recalled earlier, then this claim may buttress the belief that this item was experienced originally.

To investigate such beliefs, we revisited an older literature on output monitoring that was originally devised by Gardiner and his colleagues (Gardiner & Klee, 1976; Gardiner, Passmore, Herriot, & Klee, 1977; Klee & Gardiner, 1976, 1980; Klee & Legge, 1980). In this work on output monitoring, participants were given a series of short recall or recognition tests on unrelated words, which was otherwise very similar to the false memory procedure reported by Roediger and McDermott (1995). After the last test in the series, participants were provided with all of the studied items for the output monitoring test. This

task required participants to identify by a forced-choice, yes-no judgement whether or not they had previously recalled (or recognised) each item. In this way, people's memory for their past memory performance was assessed.

As applied to the false memory paradigm, Gardiner's technique introduces three distinct phases of the experiment: study, recall, and test of past recall performance. The beliefs we are testing in this study concern recall of critical lures earlier in the second stage as measured in the third stage of the experiment. Participants could claim that they had or had not recalled items that either they did or did not recall. In this fashion, the output monitoring test is not directly assessing participant's beliefs about what they experienced during encoding *per se*, but rather, it is assessing what they remember having written down during the free recall test. Because it is reasonable to assume that participants who claim to have recalled an item also believe it was studied originally, the output monitoring test would allow us to assess beliefs about critical lures that either did or did not constitute false memories during the earlier free recall period. Whereas output monitoring measures memory for whether items were recalled, input monitoring measures memory for whether items were studied during encoding. We conducted such a combined input-output monitoring test after six lists of Roediger and McDermott's (1995) associates were studied and recalled. On the final combined input-output monitoring test, we asked participants to identify old and new items as new (neither studied nor recalled), as studied and not recalled, or as studied and recalled during the free recall tests.

Theoretically, our primary interest was in studying memory for the critical lures that were inserted into the final input-output monitoring test. When a critical lure was erroneously recalled, we expected that participants would identify it as a studied and recalled item (hereafter, abbreviated as an SR item). If that belief persists (which it should), then the item should be labelled SR on the final input-output monitoring test. However, when a critical lure was not erroneously recalled on the earlier free recall test, then the appropriate response would be to label it new on the input-output monitoring test (i.e., not studied and not recalled). However, to the extent that participants label an unrecalled critical lure as an SR item, that claim would represent evidence for a dual false memory effect. We say dual effect because critical lures never appeared at encoding, and if they were

not recalled but participants claim that they were recalled, then such claims are indicative of a false memory (that it was recalled) of a false memory (that it was originally studied). We expected attributions for critical lures that were not initially recalled would be for participants either to label them new, or studied and not recalled (hereafter, abbreviated as an SNR item). The important point is that the three-alternative test format of the final input–output monitoring test (i.e., SR, SNR, and new) will allow us to compare beliefs concerning the critical lures that were recalled versus those that were not. Obviously, entirely new items that were not semantically related to the list items should largely be labelled new.

The current experiments were not designed to provide direct tests of competing theories for the occurrence of false memories. However, two of the widely cited mechanisms include appeals to fuzzy trace theory and implicit associative responses (see Robinson & Roediger, 1997). Fuzzy trace theory proposes that dual memory codes are established at encoding (e.g., Brainerd, Reyna, & Brandse, 1995; Reyna & Lloyd, 1997). These include verbatim traces and semantic gist traces. False memories tend to occur under those conditions when participants rely on gist during a memory test rather than verbatim memories. By contrast, implicit associative responses (IARs) are automatic associations that are generated in response to the semantic meaning of list items at encoding, although they could presumably occur on a recognition test if sufficient numbers of list items are present. Nevertheless, the results from the present study could inform these theories. For example, if the results demonstrated that participants claimed to study and to recall a critical item they did not recall earlier, then some would take this as evidence in support of greater reliance on gist memory after a delay, and hence evidence in favour of the mechanisms inherent to fuzzy trace theory. Of course, proponents of IAR theory could argue that the false memory was evoked during recall but was withheld, or that the IAR occurred during the output monitoring test itself. We will test some conditions that are relevant to such predictions, but we will keep our claims conservative as to how well our results constrain one theoretical alternative versus the other.

By way of overview, we conducted three experiments. In the first, input–output monitoring claims were tested for critical lures that had and had not been recalled. That test was conducted in two ways: one with items from each of the lists

blocked at test and one in which the items from all lists were randomly intermingled together. This blocking manipulation has increased the incidence of false recognition when lists were blocked as compared with random presentation at encoding (Mather et al., 1997). We used the blocked version of the test to assess whether IARs might occur during testing. If they do, then erroneous claims should be higher under blocked as opposed to mixed testing. In this experiment, we used Gardiner et al.'s (1977) technique of having the free recall list absent (i.e., words disappeared after they were typed into the computer). Because rates of false memories were very high, an attempt was made in Experiment 2 to reduce the erroneous claims on the final test by leaving the recall protocol on the computer during the free recall tests. In Experiment 3, the response options were changed in another attempt to reduce the high levels of erroneous claims that were exhibited on the input–output monitoring test.

EXPERIMENT 1

In Experiment 1, the six lists from Roediger and McDermott (1995) were used that produced the highest incidence of false recall. Items from a given list were presented one at a time and after a short, filled retention interval a free recall test was given. After all six lists had been presented and tested in this fashion, the final input–output monitoring test was administered, in which participants made claims about whether items were studied and recalled (SR), studied but not recalled (SNR), or were new. In two separate conditions, the test items (studied items, critical lures, and new items) were either blocked by list or randomly intermingled with items from other lists. As discussed earlier, critical lures that had been recalled should be labelled SR. One key question involves what participants would claim concerning the critical lures that went unrecalled. If the discrimination on the input–output monitoring test requires cognitive processing similar to source attributions, then unrecalled critical lures should not possess characteristics consistent with either the encoding session or the free recall test. Therefore, lacking attributes from either experience, they should be labelled new. However, to the extent that unrecalled lures come to possess attributes of list items (e.g., Hicks & Marsh, *in press*), these items may be claimed to have been studied but not recalled (SNR). To the extent that

any substantial proportion of the unrecalled lures are subsequently attributed to past experience on the input–output monitoring test, the balance of SNR versus SR claims will be revealing about changes in the phenomenological experience that occur from the free recall test in the second stage of the experiment to the input–output monitoring test in the third stage.

Method

Participants. In exchange for partial credit towards a course research requirement, 60 undergraduates from the University of Georgia volunteered. Each participant was tested individually. Thirty participants were randomly assigned to the blocked testing condition in which words from a given studied list were intermingled with new items and the critical lure, and then presented together on successive trials of the input–output monitoring test. The remaining participants received a mixed test in which all test items were randomly mixed together.

Materials and procedure. The *sleep*, *needle*, *sweet*, *mountain*, *chair*, and *rough* lists from Roediger and McDermott's (1995) appendix were presented at encoding. Each list comprised 15 items. Software was written to present items at a 2-second study rate with a 500 ms inter-stimulus interval. Following presentation of a given list, the computer presented a series of double-digit addition problems (randomly generated) for which the participant was asked to type in the answer. This arithmetic distractor period lasted a total of 30 seconds, and to maintain attention, feedback was provided as to whether answers were correct or not. A 1-minute recall period ensued in which participants were instructed to free recall the list. Participants typed the recalled words into the computer and after pressing the ENTER key for each word it was erased. Thus, as in Gardiner et al.'s (1977) procedure, recall feedback was impaired because participants could not externally review the words that they had already recalled. This procedure was repeated in an identical fashion for all six lists. The presentation order of the lists was randomly determined anew for each participant.

Following the six study and free recall segments, the input–output monitoring test was administered (i.e., approximately 12 minutes after the experiment commenced). This test differed

depending on whether participants were assigned to the blocked versus mixed condition. In both conditions, 32 trials per list were tested (for a total test of 192 trials). In the blocked condition, the 15 studied items were randomly intermingled anew for each participant with the nonpresented critical lure and 16 new items that were semantically dissimilar from any studied item. The list that was studied first was then tested first, and the list that was studied second was then tested second, and so forth, in succession. The test was self-paced and the software paused briefly between lists (i.e., after every 32 test trials). The inter-stimulus interval was 500 ms. Participants had been instructed that they were to respond by pressing one of three labelled keys to indicate whether the item was studied and recalled (SR), was studied and not recalled (SNR), or was new. The procedure in the mixed condition was similar except that all of the items on the test were randomly intermingled with each other. Thus, the 90 studied items from the 6 lists, the 96 new words, and the 6 critical unrepresented lures appeared randomly under no particular constraints for a given participant. The software also paused briefly (as in the blocked condition) after every 32 items.

Results and discussion

Unless otherwise noted with a specific p value or a statement of non-significance, all statistical tests are significant at the conventional .05 alpha level. The data from the free recall tests and a portion of the data from the combined input–output monitoring test are summarised in the upper third of Table 1. Rows of the table correspond to different aspects of performance that will be discussed sequentially.

Following the presentation of each list, free recall was assessed immediately after the 30-second distractor period. As seen in the first row of Table 1, the blocked versus mixed testing which manipulated the characteristic of the subsequent input–output monitoring test did not affect the overall proportion of studied items that were recalled. Thus, there was no sampling bias in assignment of participants to conditions, $t(58) < 1.0$, n.s. There were slightly more intrusions of the critical lures in the blocked than in the mixed condition but this difference was not statistically significant, $t(58) = 1.48$, $p > .15$. The overall levels of correct recall and critical intrusions are quite similar to previous reports (e.g., Payne et al., 1996;

TABLE 1
Performance expressed as proportions

Dependent Measure	Blocked		Mixed	
	M	SE	M	SE
<i>Experiment 1</i>				
Free Recall				
Correct Recall	.57	.01	.59	.02
Critical Intrusions	.54	.04	.45	.04
Input–Output Monitoring				
Inferred Hit Rate	.86	.01	.86	.01
Unrelated FA Rate	.11	.01	.09	.01
Critical Lure FA Rate	.94	.02	.89	.03
<i>Experiment 2</i>				
Free Recall				
Correct Recall	.58	.02	.56	.02
Critical Intrusions	.48	.04	.52	.04
Input–Output Monitoring				
Inferred Hit Rate	.87	.01	.85	.02
Unrelated FA Rate	.11	.01	.12	.01
Critical Lure FA Rate	.89	.03	.87	.04
<i>Experiment 3</i>				
Free Recall				
Correct Recall	.63	.02	.61	.02
Critical Intrusions	.52	.05	.37	.05
Input–Output Monitoring				
Inferred Hit Rate	.87	.02	.84	.02
Unrelated FA Rate	.11	.01	.10	.01
Critical Lure FA Rate	.90	.03	.74	.04

Performance expressed as proportions for Experiments 1–3 on the free recall tests and the inferred recognition hit rates from the input–output monitoring tests.

FA = False Alarm, M = Mean, SE = Standard Error

Roediger & McDermott, 1995). One important aspect of performance is that the incidence of critical intrusions hovers around 50%. This rate is ideal for making comparisons between critical intrusions that were recalled versus those that were not as tested in the subsequent input–output monitoring test because approximately equal numbers of each were obtained. Numbers of extra-list intrusions and recall repetitions were quite miniscule as is typical in this paradigm and they will not be considered further (for a discussion concerning their insignificance see Robinson & Roediger, 1997).

For performance on the input–output monitoring test, a measure of the inferred recognition hit rate combines the proportions of studied items labelled SR and SNR (i.e., the probability of calling an item “studied” regardless of the claim of whether it had been recalled). These proportions are also reported in Table 1. Performance on old items was quite good at 86% for both conditions. Testing all of the list items together in

blocks might have supported better identification than randomly mixing all lists together, but it did not, $t(58) < 1.0$, n.s. Thus, unlike previous manipulations in which this variable had an effect at encoding (e.g., Mather et al., 1997) it did not appear to affect performance when introduced at test. The new items were semantically unrelated to list items and should have been easily rejected. Although a small proportion of them were claimed to be old (approximately 10%), this false alarm rate did not differ by condition tested, $t(58) < 1.0$, n.s. One interesting aspect of performance was that participants claimed the critical lure was on the study list approximately 90% of the time. Given that veridical recognition of the studied items was only 86%, these data suggest that participants believed that critical lures were experienced at a rate slightly higher than truly studied items. Although the difference was small, in a 2×2 mixed model ANOVA with condition (blocked versus mixed) and item type (studied items versus critical lures), the effect of item type was statistically significant, $F(1,58) = 6.62$.

We turn now to participants’ assertions about whether or not they had recalled the critical lures earlier during the free recall test. These data for the critical lures *only* are summarised in Table 2. In that table, performance is reported and analysed separately for the critical lures that were recalled and those that were not recalled earlier. Therefore, for each type of critical lure (recalled versus not recalled) the proportions in the two columns of means sum to 1.0. Because one or two subjects in each condition either recalled all of the critical lures or recalled none, degrees of freedom are reduced slightly and will differ accordingly. Finally, it is impossible to conduct ANOVAs on proportions that sum to unity, and therefore, we rely on simple between-conditions comparisons and common-sense examination of the relevant cell means (for which standard errors have been provided as well).

For the critical lures that were erroneously recalled, performance was similar in the blocked and mixed conditions, all three $t(57) < 1.0$, n.s. However, the claims for these items were far from equivalent. Participants generally remembered recalling the critical lures, and therefore asserted that they were both studied and recalled. As noted in the introduction, this outcome was expected if participants’ beliefs remained stable from free recall to the input–output monitoring test. There was some forgetting that the critical lures had been recalled because approximately 20% of

TABLE 2
Claims about critical lures

<i>Recall Status and Claim</i>	<i>Blocked</i>		<i>Mixed</i>	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
<i>Experiment 1</i>				
Critical Lures Recalled				
New	.02	.01	.01	.01
Studied and Not Recalled	.17	.05	.23	.06
Studied and Recalled	.81	.05	.76	.06
Critical Lures Not Recalled				
New	.08	.03	.15	.04
Studied and Not Recalled	.44	.06	.46	.06
Studied and Recalled	.49	.07	.39	.06
<i>Experiment 2</i>				
Critical Lures Recalled				
New	.01	.01	.07	.03
Studied and Not Recalled	.12	.05	.08	.03
Studied and Recalled	.87	.05	.85	.04
Critical Lures Not Recalled				
New	.18	.04	.17	.05
Studied and Not Recalled	.39	.06	.40	.06
Studied and Recalled	.43	.06	.43	.07
<i>Experiment 3</i>				
	<i>Absent</i>		<i>Present</i>	
Critical Lures Recalled				
New and Not Recalled	.03	.02	.02	.02
New and Recalled	.00	.00	.11	.05
Studied and Not Recalled	.09	.05	.02	.02
Studied and Recalled	.88	.05	.84	.06
Critical Lures Not Recalled				
New and Not Recalled	.17	.06	.26	.06
New and Recalled	.06	.04	.05	.02
Studied and Not Recalled	.22	.06	.35	.06
Studied and Recalled	.55	.08	.34	.06

Claims in the input–output monitoring test about the critical lures that had and had not been falsely recalled during free recall.

M = Mean, SE = Standard Error

these items were labelled studied and not recalled. The data of most interest concern the claims for critical lures that had not been recalled during the free recall test. For these unrecalled lures, responses were not uniform across the three options, although blocked versus mixed testing did not influence performance, all three $t(57) < 1.0$, n.s. Surprisingly, very few of these items were labelled new as they should have been (only about 11%). Rather, participants claimed that they were either studied and not recalled (SNR) or that they were studied and recalled (SR). This belief of studying and recalling (when neither event actually occurred) was slightly larger for the blocked condition (49%) than the mixed condition (39%) but the two conditions do not differ statistically as can be inferred from the standard errors. Recall that approximately 50% of the

critical lures were recalled and the overwhelming majority of those were labelled SR. Therefore, about half the time when an unrecalled critical lure was claimed to have been studied, participants believed that they had also recalled it earlier when in fact they had not. Performance on this input–output monitoring test suggests that one can come to believe erroneously that an item was studied (when it was not) but also that one had actually recalled it (when one had not). This outcome clearly represents a dual false memory effect for unrecalled lures (i.e., a false memory of a false memory).

The failure of the mixed versus blocked testing conditions to produce significant differences suggests that IARs are not occurring at different rates in the two types of test. Therefore, this null outcome lends some credence to the notion that IARs, if they are occurring at all, must occur during the original study episode. By the same token, the null effect of mixed versus blocked testing also suggests that if participants are relying on gist representations during the input–output monitoring test, then a blocking manipulation does not increase this reliance on gist representations.

A 3×3 table of claims could be provided for the studied items in which the state of nature (SR, SNR, and new) of these items from the free recall test is compared to participants' claims on the input–output monitoring test (SR, SNR, and new). However, for the sake of brevity we simply report the highlights and the important comparisons between the critical lures and the studied items. Basically, on the input–output monitoring test people were quite accurate at identifying the unrelated lures as new (90%) and the items they had studied and recalled as such (81%). Thus, critical lures that had been erroneously recalled were labelled SR at the same rate that studied items were recalled and labelled SR. Studied items that were forgotten were correctly labelled SNR only 55% of the time. Perhaps because they went unrecalled, these items were labelled new 30% of the time. Therefore, the remaining 15% of these SNR items were labelled studied and recalled when in fact they were not recalled. The relevant comparison to be made is that in Table 2, 44% of the unrecalled critical lures were labelled studied and recalled (averaging over the blocked and mixed conditions) whereas participants made this erroneous claim of false recall for nonrecalled studied items only 15% of the time $t(58) = 6.0$. Therefore, the characteristics of critical lures are

quite special in creating beliefs that they had been recalled when in fact they were neither studied nor recalled.

By way of summary, participants were quite accurate at identifying that they had recalled a critical lure, but as a consequence, they were claiming that it was studied (which it was not) and recalled (which it erroneously was). In other words, beliefs were stable about false memories that had been created earlier. In addition, new false memories were created in so far as participants asserted that unrecalled critical lures were studied about 90% of the time and further that they had actually been typed into the computer approximately half of that time. This last effect is a dual error representing false memories (that an item was recalled) of false memories (that an item was studied). To the extent that critical lures came to mind and were rejected during recall, this effect also represents a substantial change in beliefs from the recall test to the final test. Toglia, Neuschatz, and Goodwin (1999) demonstrated that over a delay in this paradigm recall for studied items declines whereas recall for critical lures remains stable. We have now found an increase in claims concerning the critical lures, and that result seems to be more consistent with participants basing their judgements on a gist representation as is asserted in fuzzy trace theory.

EXPERIMENT 2

The input-output monitoring test following free recall of each list is one measure of participants' beliefs about their free recall performance. We observed that if participants had a false memory of the critical lures during recall, those false beliefs persisted. We thought that the additional cognitive processing that must be brought to bear in discriminating between memory for input versus output (as compared to recognition) should cause memories to be inspected in a more detailed fashion. In turn, more stringent decision criteria should have created some caution in assessing what was studied versus what was recalled. Contrary to these expectations, almost every critical lure that went unrecalled was labelled studied, and half of those were falsely believed to have been recalled (but actually were not). The creation of these new, dual false memories may have been exacerbated by participants' impaired memories of the recall episodes, because each recalled word

was immediately erased after it was typed into the computer. In Experiment 2, the recall protocol was left on the monitor in order to ascertain if the additional encoding during recall output would reduce the false claims of having recalled critical lures. Note that this extra processing is occurring during the recall test itself (i.e., in the second phase of the experimental sequence) and is predicted to affect performance during the input-output monitoring phase of the experiment (i.e., the third phase). Therefore, we expected better output monitoring and better input monitoring. Although blocked versus mixed testing did not affect claims about memory in Experiment 1 this variable was manipulated again to ascertain whether it might have an effect with potentially better performance when the recall protocol was present.

Method

Participants. For course credit towards fulfilling a research requirement, 60 undergraduates who had not participated in Experiment 1 were recruited from the same pool. Although half of the participants were to be assigned to blocked versus mixed testing, the misassignment of one participant resulted in 31 participants being tested in the blocked condition and 29 being tested in the mixed condition.

Materials and procedure. The materials and procedure were virtually identical in all respects to Experiment 1. The only difference concerned the free recall tests. After studying a list of associated words and doing double-digit arithmetic problems, free recall was recorded for 1 minute as described in Experiment 1. In this experiment, participants typed in words on the computer, starting their protocol in the upper left-hand corner of the blank screen. After pressing the ENTER key, the word remained on the screen and the cursor moved to the next line for the subsequent word to be recalled. This procedure was repeated for each word such that the recall protocol was recorded in a column down the left-hand side of the computer monitor that participants were free to review during the recall period. After 1 minute the computer beeped and informed the participant to prepare for studying the next list. If the participant was in the middle of typing a word at the deadline, then the computer allowed the participant to finish typing it.

Results and discussion

As will quickly become evident, the results of this experiment confirmed those of Experiment 1 in their entirety (see the middle third of Table 1). The proportion of words correctly recalled was close to 60% and did not differ for the mixed versus blocked testing conditions, $t(58) < 1.0$, n.s. In fact, none of the differences between the blocked and mixed conditions was significant for any of the dependent measures in Tables 1 or 2, and therefore, statistical analyses between these conditions will be omitted for the sake of brevity. This outcome bolsters the interpretation that IARs do not occur during testing, because if they did then they should have been more prevalent in the blocked condition. As before, the proportion of critical intrusions on the free recall test was approximately 50%. This interim level of performance is ideal for comparing beliefs about critical intrusions that were recalled previously versus those that were not.

The inferred hit rate on the input–output monitoring test combines claims that items were SR and SNR (i.e., that items were studied). As in Experiment 1, this rate was approximately 86%. Although the false alarm rate for critical lures was statistically higher than the hit rate for studied items in Experiment 1, making the recall protocol available in this experiment equated these two rates. In the 2×2 ANOVA with item type (critical lures versus studied items) and condition (blocked versus mixed), there was no effect of item type, $F(1,58) < 1.0$, n.s. As before, the false alarm rate for semantically unrelated items was small.

Participants' claims about their free recall of the critical lures are summarised in the middle third of Table 2 in an identical manner as described before. The null effects of blocked versus mixed testing as well as the equivalent results obtained with impaired versus unimpaired free recall across the two experiments adds some generality to the outcomes reported in Table 2. In other words, the input–output monitoring test appears to be a very stable measure of participants' beliefs concerning their past performance because the manner in which the test was administered did not affect performance. (For data and discussion of how source judgements can depend on the manner in which a test is administered, see Marsh & Hicks, 1998.) For the critical lures that were recalled during the free recall test, the majority of these were identified as SR, which

replicated Experiment 1. Therefore, participants largely remembered that they had recalled them, indicating that false beliefs of having studied them persisted.

Claims that unrecalled critical lures were indeed new items rose slightly from Experiment 1 as a result of making the recall protocol available in this experiment (from approximately 12% to 18%). This outcome is consistent with better verbatim memory for the recall episode. Nevertheless, over 80% of the critical lures that were not recalled were claimed to have been studied when tested in the input–output monitoring test (i.e., adding together the claims of SR and SNR), and about half of those were believed to have been recalled during the free recall test (i.e., SR). As compared to previous investigations of false memories, these claims are novel in so far as they represent a belief that items were studied (when they were not) and also a belief that they were typed into the computer during the free recall test (which they were not). As argued before, the input–output monitoring test has revealed a double-layer false memory effect. More importantly, that effect represents a false claim of having experienced an item on more than one occasion. As we argued earlier, these results appear to be more consistent with a fuzzy trace explanation of false memories, because IAR theory would have difficulty explaining why people come to believe that unrecalled lures were indeed recalled. By contrast, gist-based reliance at test handles this result more easily.

Similar to the previous experiment, people identified the majority of new items in the input–output monitoring test as new items (89%) and they largely correctly identified the items that they had studied and recalled as such (86%). The items that were truly studied and not recalled were either attributed correctly as SNR items (57%) or were incorrectly called new (28%). These claims left 15% of the studied items that went unrecalled (i.e., SNR items) erroneously labelled studied and recalled. That 15% SR claim for unrecalled studied items is far smaller than the 43% SR claim in Table 2 for unrecalled critical lures, $t(57) = 6.2$. In other words, unrecalled critical lures are identified as SR items much more often than unrecalled studied items, despite the fact that neither the study event nor the recall event actually occurred for critical lures whereas studied items were at least seen at encoding. This claim is a somewhat more powerful and compelling memory distortion than simply claiming that an item was seen for 2 seconds at encoding.

EXPERIMENT 3

The obvious conclusion from Experiments 1 and 2 is that the blocked versus mixed manipulation does not affect performance on the combined input–output monitoring test. Therefore, this manipulation was abandoned. Although facilitating versus impairing free recall did not have dramatic effects when tested across Experiments 1 and 2, its manipulation as two conditions within one experiment should provide greater resolution to detect any potential effect on participants' assertions concerning their false memories. In other words, cross-experimental comparisons across Experiments 1 and 2 are vastly inferior to manipulating that variable within the same experiment. Therefore, in two between-subjects conditions we tested leaving the recall protocol on the screen (cf. Experiment 2) versus impairing output monitoring by immediately erasing each recalled word (cf. Experiment 1). Protocol presence should reduce false recall because having the protocol available will decrease the memory load for deciding which items had and had not yet been recalled during the free recall test. In turn, with a lighter cognitive load candidate memories may be inspected more deeply and critical lures edited out more efficiently.

We also took the opportunity in Experiment 3 to address a potential concern of the first two experiments. Thus far the input–output monitoring test has been conducted using three options (SR, SNR, or new). We found very high rates of believing that critical lures were studied and recalled, when in fact, they were never studied or recalled. This degree of memory distortion could be a consequence of a slightly unbalanced set of test options. Two of the options (SR and SNR) are claims of being old whereas only one option is a claim of being new. Donaldson, MacKenzie, and Underhill (1996) have argued that the use of such unbalanced scales on recognition tests induces liberal responding (on their scale one option is new and three options are old).

To ascertain if a similar cognitive bias occurred during the input–output monitoring test in the first two experiments, participants in Experiment 3 were given two responses to claim that a critical lure was old (SR and SNR). They were also given two options with which they could claim it was new. These two options included the claim that an item was new and it was not recalled previously (the identical option used before). The second new option was the claim that the item was new

but the participant believed he or she had erroneously recalled it during the free recall test. By changing the response options in this fashion we were able to determine if participants performing the input–output monitoring test realised that they had accidentally recalled a critical lure that was not presented in the study list. In addition, the added option should increase deliberations concerning the status of items, thereby further increasing the need to consider the source specifying information associated with the items.

Method

Participants. A total of 52 University of Georgia undergraduates volunteered in exchange for partial credit towards a course requirement. None had participated in Experiments 1 or 2. Participants were tested individually. Of the volunteers, 26 were assigned to the recall absent condition and the remaining 26 participants were tested with their recall protocol present on the screen.

Materials and procedure. The materials and basic procedure were virtually identical to those used before. The input–output monitoring test was blocked by list. In the protocol absent condition, recalled words were erased immediately after they were typed in, just as in Experiment 1. In the protocol present condition, the recalled words appeared down the left-hand side of the monitor, just as in Experiment 2. To maximise the probability of finding differences between these two conditions the free recall period was increased from 1 minute to 1.5 minutes in this experiment. Because the majority of the 120 participants tested in the previous two experiments had finished their recall by the end of 1 minute, we hypothesised that the additional 30 seconds might be used fruitfully in studying the recall protocol. This additional exposure during free recall should result in information that could be used to avoid erroneously identifying the critical lures as recalled, because deeper or better encoding generally leads to better source monitoring (e.g., Johnson, Hashtroudi, & Lindsay, 1993).

The only other procedural change involved the number of options on the input–output monitoring test. Previously, to identify an item as old participants could claim that an item was studied and recalled (SR) or an item was studied and not recalled (SNR). These options remained as

before. In this experiment, two new options were available to balance the number of new and old alternatives. Participants were instructed that for new items they could claim the item was not presented and they had not recalled it, or they could claim that an item was new (i.e., unstudied) and that they had erroneously recalled it during the free recall test.

Results and discussion

Performance is summarised in the lower third of Table 1 in an identical manner as reported previously. Free recall did not covary with whether the protocol was present or absent, $t(50) < 1.0$, n.s. The incidence of false recall was approximately 50% in the protocol absent condition but was significantly less (by 15%) in the protocol present condition, $t(52) = 2.14$. This result suggests that allowing participants to review their recall protocol for a long enough period of time can aid in reducing false recall by minimising the recall burden involved with discriminating what had been recalled from what had not yet been recalled.

The inferred hit rate did not differ as a function of the recall protocol being present or absent, $t(50) = 1.15$, n.s. Claims that the critical lures were studied were entered into a 2 (item type: critical lure versus studied item) \times 2 (condition: protocol absent versus present) ANOVA similar to the previous experiments. The interaction term was statistically significant, $F(1,50) = 6.41$. This interaction indicated that making the recall protocol available on the screen significantly reduced the erroneous beliefs that critical lures were studied (i.e., false memories) but did not affect claims concerning studied items. In light of these fewer intrusions, that result strongly suggests that later false memories can be reduced by having the recall protocol available for participants to consult if the duration of the recall period is sufficiently long. Stated alternatively, this result suggests that the incidence of false memories is greater when previous recall does not leave a permanent record that can be inspected. Obviously, testing the protocol absent versus present conditions within this experiment produced much greater effects than inferring its manipulation across experiments. Moreover, providing a lengthened recall period also allowed participants to study their recall protocols and this extra time resulted in a reduction of false memories. Together all of these results are consistent with participants relying

more on verbatim memories in the protocol present condition as compared with the absent condition.

Participants' claims during the input-output monitoring test about whether critical lures were studied and recalled are summarised at the bottom of Table 2. The vast majority of critical lures that were erroneously free recalled were labelled studied (which they were not) and recalled (which they were). This finding nicely replicated the results of both Experiments 1 and 2. However, having the protocol available resulted in participants claiming that they had erroneously recalled some critical lures during the free recall test at least a small proportion of the time (11%), whereas those assigned to the protocol absent condition believed they had studied them but failed to recall them (9%). Thus, these data indicate (consistently with Experiments 1 and 2) that once a critical lure is recalled, participants' beliefs generally reflect that it was both studied and recalled. The fact that some small proportion of false memories can be identified as such in the final test is extremely important because it suggests that under some circumstances participants can identify their false memories as just that, a false memory representing an erroneous *earlier* belief. As discussed later, the challenge facing researchers will be to find conditions where that 11% identification rate approaches much higher levels.

For critical lures that were not recalled during the free recall tests, fewer items were identified as SNR and more were claimed to be SR pooling over assigned condition. Of critical lures believed to be studied (SR plus SNR), approximately equal numbers were asserted to have been recalled versus not recalled in the protocol present condition. By contrast, in the protocol absent condition more critical lures were labelled recalled than not recalled. This finding is supported by a significant (21%) difference in SR items across the present versus absent conditions, $t(48) = 2.10$. Furthermore, more items were claimed to be new and not recalled when the recall protocol was present than when it was absent. Together, these last two results suggest much better input and output monitoring when the recall protocol was available and extra time to study it was provided (i.e., a longer recall period).

The more dramatic outcome of this experiment was that with a 1.5-minute recall period with the protocol present, half of the unrecalled critical lures were believed to be SR with the remaining

half being labelled only studied (SNR). This 50–50 pattern is to be compared with removing the recall protocol which resulted in about a 70–30 split belief that the unrecalled lures were predominantly both studied and recalled (when neither event truly occurred). Thus, protocol absence appeared to exacerbate the belief that items were recalled when in fact they were not. As noted earlier, these distortions of memory are powerful because participants are asserting not only having encountered items on the study list but also claiming to have a memory of typing it into the computer, when in fact, neither event actually occurred.

GENERAL DISCUSSION

These experiments were conducted to explore what people believed concerning the false memories that were created by (and those that lay dormant during) a free recall test. In general, erroneous beliefs that critical lures were experienced will persist if they have been mistakenly recalled earlier. After all, over 80% of these items were labelled as studied and recalled across a variety of conditions. By contrast, unrecalled critical lures were also asserted to have been studied over 80% of the time. The finding that recalled and unrecalled lures were claimed to have been studied at similar rates is intriguing because the latter items did not pass any subjectively set criterion for outputting items during the free recall test (or simply were not generated as candidate memories during recall). Therefore, although these two classes of items did not behave similarly during the free recall memory test, claims of having been studied were nevertheless similar during the subsequent input–output monitoring test. The novel finding was that 50% of the unrecalled lures were asserted to have been recalled, which meant that participants believed that they had typed them into the computer. This mistaken claim of studying and recalling the unrecalled lures has been labelled a false memory of a false memory because there is some duality about participants claiming that they had studied and recalled an item that was never presented and never recalled.

These effects might have been larger with blocked rather than random testing, but ultimately that manipulation did not affect performance. Nevertheless, we have demonstrated that the particular alternatives used on the input–out-

put monitoring test can influence the claims that are made. In Experiment 3, a more balanced set of options was adopted including one that allowed participants to claim that they had mistakenly recalled the critical lure before. People did use this option, but only very infrequently (at an 11% rate) and only when the recall protocol was available for people to study during free recall in the second stage of the experiment. By an IAR hypothesis, testing all of the items from the same list together could have evoked IARs at test in a similar manner to encoding. This outcome did not occur and lends some credence to the idea that false memories may occur because participants are relying on gist representations to make their decisions.

At several points in this article we have argued that erroneous claims of recalling represent a more powerful memory distortion as compared to claims that an item was only studied. We believe that this is particularly true given that discriminations on the input–output monitoring test should have required fairly detailed inspection of memories akin to source judgements. However, false memories may occur in the first place as a consequence of the associative nature and structure of items encountered during the encoding experience (see Robinson & Roediger, 1997). In this light, our finding that people believed they had recalled an unrecalled item may be the consequence of an inference. Because of semantic dissimilarity, the discrimination between studied versus unstudied items should have been relatively easy. The greater difficulty for people was determining whether an item was recalled or not. Having identified an item as studied, discrimination processes may not locate definitive attributes that indicate its status as recalled versus not. Given that over 50% of the studied items were recalled, people may judge by inference that some proportion of the unrecalled lures were recalled. This argument is entirely consistent with gist-based responding as argued by fuzzy trace theory.

One promising method to titrate beliefs of having studied versus recalled would be to adopt Robinson and Roediger's (1997) manipulations of progressively watering down the association value of the list with semantically dissimilar extra-list items. For example, as the association value of the list is gradually weakened with more unrelated words at encoding, erroneous claims of having studied critical lures may decline at less precipitous rates as compared to a more rapid decline in the mistaken claims of having recalled an item.

Of course, such experiments might prove to be more informative if they were tested with the four-alternative test options introduced in Experiment 3. Because such manipulations make truly new items in the input–output monitoring task less differentiated from extra-list items, the final test would be more challenging and therefore potentially require people to adopt even more cautious criteria about their claims (because new and old items would be less differentiated from one another).

These predictions are based on our own belief that tasks and conditions could be found that will allow people to identify their mistakes of having recalled a critical lure that was never presented. Although our own beliefs may be as erroneous as the beliefs that were instilled in our participants, we have noted the strong similarity between the requirements of the input–output monitoring task and making a source judgement (i.e., discriminating the claims of SR, SNR, and new). In such explicit tests, we argued that source judgements can be cognitively demanding because the alternatives need to be weighed carefully and criteria must be established to choose among the various alternatives. From this perspective, participants should have been able to identify unrecalled critical lures as new items. Because ultimately they could not, it may be unreasonable for us to claim that participants could be made to edit out these intrusions to any efficient extent.

Consistent with the idea that it will be difficult to find conditions that efficiently inoculate people from false memories in this paradigm, Roediger and McDermott (1995) demonstrated that a majority of them were labelled “remember” rather than “know” using the remember–know paradigm (e.g., Gardiner & Gregg, 1997; Gardiner & Java, 1990). Likewise, Mather et al. (1997) demonstrated that people were willing to ascribe perceptual detail from the encoding experience to their false memories and otherwise claim that they were experienced from a particular source (see also Hicks & Marsh, 1999, Norman & Schacter, 1997; Payne et al., 1996). In addition, warnings about the effect will reduce but never entirely eliminate false memories (Gallo et al., 1997; McDermott & Roediger, 1998). These findings suggest that the encoding experience can evoke these false memories very powerfully (cf. Robinson & Roediger, 1997). In addition, the present results may demonstrate that false beliefs can rise over time, thereby becoming more prevalent as compared to veridical memory for studied items

(cf. McDermott, 1996; Toggia et al., 1999). Obviously, this prediction could be tested by delaying the final input–output monitoring test across several different retention intervals and examining people’s beliefs about their previous recall using the procedure we have adopted.

To assess beliefs about past performance, we have reintroduced an alternative method that can substitute for taking confidence judgements or using the remember–know paradigm (at least in some situations). In Roediger and McDermott’s (1995) false memory paradigm, we found an incredibly robust belief that critical lures were studied, but more importantly that they had often been recalled. We are not sure at what point in the experimental sequence the belief was created that an unrecalled item was recalled. Because input and output monitoring was tested after free recall, the mistaken beliefs that items were recalled might arise from considering critical lures as to-be-recalled candidates during free recall, or those beliefs may have been created at encoding as Robinson and Roediger (1997) have claimed (see also Roediger, McDermott, & Robinson, 1998). One way to examine this issue would be to change the nature of the free recall test. For example, recall could be cued with fragments of studied words and/or the number of fragments of studied items per list could be manipulated as a means of examining how this variable would affect erroneous claims of having recalled an unrecalled item. Whatever the results of the several Gedanken experiments we have proposed in this discussion section, we hope that others will view this paradigm as a useful extension to previous manipulations and use it to explore the theoretical underpinnings of the false memory phenomenon. If nothing else, these experiments demonstrate that older paradigms can be recycled and put to exceptionally good use to explore the current problems and puzzles that continue to generate excitement and interest.

Manuscript received 2 March 2000

Manuscript accepted 31 August 2000

REFERENCES

- Bower, G.H., & Mann, T. (1992). Improving recall by recoding interfering material at the time of retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1310–1320.
- Brainerd, C.J., Reyna, V.F., & Brandse, E. (1995). Are children’s false memories more persistent than

- their true memories? *Psychological Science*, 6, 359–364.
- Deese, J. (1959). On the prediction of occurrence of a particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, 58, 17–22.
- Donaldson, W., MacKenzie, T.M., & Underhill, C.F. (1996). A comparison of recollective memory and source monitoring. *Psychonomic Bulletin & Review*, 3, 486–490.
- Gallo, D.A., Roberts, M.J., & Seamon, J.G. (1997). Remembering words not presented in lists: Can we avoid creating false memories. *Psychonomic Bulletin & Review*, 4, 271–276.
- Gardiner, J.M., & Gregg, V.H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review*, 4, 474–479.
- Gardiner, J.M., & Java, R.I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, 18, 23–30.
- Gardiner, J.M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior*, 15, 227–233.
- Gardiner, J.M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior*, 16, 45–54.
- Hicks, J.L., & Marsh, R.L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1195–1209.
- Hicks, J.L., & Marsh, R.L. (in press). False identification covaries with recognition versus source tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin*, 114, 3–28.
- Klee, H., & Gardiner, J.M. (1976). Memory for remembered events: Contrasting recall and recognition. *Journal of Verbal Learning and Verbal Behavior*, 15, 471–478.
- Klee, H., & Gardiner, J.M. (1980). Remembering the recall of cued and uncued words: Effects of initial accessibility. *Canadian Journal of Psychology*, 34, 220–226.
- Klee, H., & Legge, D. (1980). Remembering the recall of words. *Canadian Journal of Psychology*, 34, 86–90.
- Lampinen, J.M., Neuschatz, J.S., & Payne, D.G. (1999). Source attributions and false memories: A test of the demand characteristics account. *Psychonomic Bulletin & Review*, 6, 130–135.
- Marsh, R.L., & Hicks, J.L. (1998). Test formats change source-monitoring decision processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1137–1151.
- Mather, M., Henkel, L.A., & Johnson, M.K. (1997). Evaluating characteristics of false memories: Remember/Know judgments and memory characteristics questionnaire compared. *Memory & Cognition*, 25, 826–837.
- McDermott, K.B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, 35, 212–230.
- McDermott, K.B., & Roediger, H.L. III. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persist under conditions of explicit warnings and immediate testing. *Journal of Memory and Language*, 39, 508–520.
- Norman, K.A., & Schacter, D.L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, 25, 838–848.
- Payne, D.G., & Blackwell, J.M. (1998). Truth in memory: Caveat emptor. In S.J. Lynn & K.M. McConkey (Eds.), *Truth and memory* (pp. 32–61). Guilford, CT: Guilford Press.
- Payne, D.G., Elie, C.J., Blackwell, J.M., & Neuschatz, J.S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, 35, 261–285.
- Read, J.D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review*, 3, 105–111.
- Reyna, V.F., & Lloyd, F. (1997). Theories of false memory in children and adults. *Learning & Individual Differences*, 9, 95–123.
- Robinson, K.J., & Roediger, H.L. III. (1997). Associative processes in false recall and false recognition. *Psychological Science*, 8, 231–237.
- Roediger, H.L. III, & McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803–814.
- Roediger, H.L. III, McDermott, K.B., & Robinson, K.J. (1998). The role of associative processes in creating false memories. In M.A. Conway, S.E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory*, vol. 2. Hove, UK: Psychology Press.
- Schacter, D.L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, 40, 1–24.
- Seamon, J.G., Luo, C.R., & Gallo, D.A. (1998). Creating false memories of words with or without list item recognition: Evidence for nonconscious processes. *Psychological Science*, 9, 20–26.
- Sommers, M.S., & Lewis, B.P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, 40, 83–108.
- Toglia, M.P., Neuschatz, J.S., Goodwin, K.A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, 7, 233–256.