INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
RICE UNIVERSITY

REMEMBERING WORDS NOT PRESENTED IN LISTS:
THE ROLE OF TESTING IN PRODUCING A MEMORY ILLUSION

by

KATHLEEN B. MCDERMOTT

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE

Henry L. Roediger, III, Professor, Director
Psychology

Jeffrey Bowers, Assistant Professor
Psychology

Stephen Klineberg, Professor
Sociology

Randi C. Martin, Professor
Psychology

Houston, Texas

May, 1996
ABSTRACT

Remembering Words Not Presented in Lists:
The role of Testing in Producing a Memory Illusion

by

Kathleen B. McDermott

The experiments reported here illuminate current understanding of the factors influencing false recall. These experiments employ a procedure reported by Roediger and McDermott (1995), who demonstrated that reliable, predictable intrusions can be elicited in single trial free recall by presenting subjects with short lists of words (e.g., bed, rest, awake, tired, dream, etc.), all of which are associates of a critical nonpresented word (e.g., sleep); the critical nonpresented word is recalled and recognized as having occurred in the list.

In Experiment 1, high levels of false recall were obtained, replicating Roediger and McDermott's (1995) results; in addition, although introduction of a short (30 s) filled delay resulted in forgetting of studied items, false recall was maintained over the delay. On a final free recall test occurring 2 days later, the probability of false recall was increased when an initial test had been taken on the first day, relative to a condition in which no initial test had occurred. In addition, the probability of false recall exceeded the probability of veridical recall on the final test. In Experiment 2, when multiple successive tests were taken, the probability that subjects would later claim to remember the critical nonpresented item (i.e., to recollect something specific about its presentation)
was enhanced relative to a single test condition.

Experiment 3 examined the persistence of this memory illusion by giving subjects multiple opportunities to hear the list (in a multtrial free recall procedure). Although the probability of free recall diminished across successive trials, robust levels of false recall were still obtained after 5 study-test trials. When subjects returned 1 day later, the probability of free recall was enhanced (relative to the last trial of Day 1), whereas the probability of veridical recall diminished across the delay.

Overall, results demonstrate that illusory memories can be extremely robust, persisting across time and repeated retrieval attempts. In some cases, the probability of false recall exceeds the probability of veridical recall. Results converge upon the conclusion that the act of retrieval can play a critical role in determining false recall.
Acknowledgments

I want to express my deepest appreciation to Roddy Roediger for his invaluable advice, unwavering support, and friendship throughout my years at Rice. I have thoroughly enjoyed working with him. His exceptional dedication to the field will continue to serve as an inspiration to me.

I would also like to thank the members of my committee (Jeff Bowers, Steve Klineberg, Randi Martin, and Roddy Roediger) for their comments and suggestions regarding these experiments. They helped me sharpen my thinking with respect to this project.

Finally, I thank Glen Doniger, Ron Haas, Bettina Johnson, Jenny Mireles, and Keith Rozendal for their help in testing subjects and coding data. Their assistance made this a manageable project and is much appreciated.
# Table of Contents

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>18</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>34</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>55</td>
</tr>
<tr>
<td>General Discussion</td>
<td>65</td>
</tr>
<tr>
<td>Footnotes</td>
<td>82</td>
</tr>
<tr>
<td>References</td>
<td>83</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Recognition results for studied items (top), critical lures (middle), and standard lures (bottom) in Roediger and McDermott's (1995) Experiment 2.  

Table 2. Proportions of studied items and critical nonstudied items recalled on the initial tests and the final free recall test as a function of initial test condition of Experiment 1.  

Table 3. Probability of recall of studied items and critical nonpresented items as a function of test number, instructional set, and time allotted for the initial tests in Experiment 2.  

Table 4. Probability of recall on the final free recall test of Experiment 2 as a function of item type, number of prior tests, and retrieval orientation on the prior tests.  

Table 5. Probability of obtaining a remember judgment on the final free recall test of Experiment 2 as a function of item type, number of prior tests, and retrieval orientation on the prior tests.  

Table 6. Testing effects in recognition memory: Recognition results from studies conforming to Roediger and McDermott's (1995, Experiment 2) procedure as a function of item type and whether an initial free recall test was taken.  

Table 7. Testing effects in Remember Responses: Results for studies
conforming to Roediger and McDermott's (1995, Experiment 2) procedure as a function of item type and whether an initial free recall test was taken.
List of Figures

Figure 1. Probability of false recall of critical nonpresented associates in Deese's (1959b) study as a function of the mean probability with which individual list items elicited the associates in a free association task.  

Figure 2. Probability of recall as a function of serial position in Roediger and McDermott's (1995) Experiment 2. (The dashed line indicates the mean probability of recall of the critical nonstudied items.)  

Figure 3. Probability of recall of studied items on the initial recall tests (both immediate and delayed) in Experiment 1 as a function of serial position of the presented items. (The dashed line indicates the mean probability of recall of the critical nonstudied items [collapsed across test condition.])  

Figure 4. Cumulative probability of recall of the critical nonpresented item as a function of output position in recall for the immediate and delayed tests of Experiment 1. Quintiles refer to the first 20% of the subject's responses, the second 20%, etc.  

Figure 5. Probability of recall of studied items on the final free recall test in Experiment 1 as a function of serial position and
initial test condition. (Dashed lines indicate probability of recall of the critical nonstudied items as a function of initial test condition.)

Figure 6. Cumulative probability of recall of critical nonpresented items and studied items recalled for each minute of the final free recall test in Experiment 1.

Figure 7. Probability of recall of studied items (solid lines) and critical nonpresented items (dashed lines) as a function of test trial and presentation order for Experiment 3.

Figure 8. Overall probability of recall of critical nonpresented items (dashed lines) and the probability of recalling “new” critical items as a function of test trial and presentation order for Experiment 3.
Introduction

The phenomenon of false memory (i.e., remembering events that never occurred or that occurred differently from the way they are remembered) is a fascinating topic that has received considerable attention in the psychological literature. As early as the 1930's, Bartlett (1932) reported that memory distortions can occur quite readily. Nevertheless, despite years of study, the phenomenon remains quite controversial and is not fully understood. The aim of the experiments reported here is to shed light on this topic by approaching questions with a new (or at least undeveloped) paradigm.

General Framework for Classifying Studies of False Memory

I begin with a very brief overview of the various ways in which researchers have attempted to study false memory effects in the laboratory. The overview is far from exhaustive; the intent is to provide a general framework so that the experiments reported here can be understood in the context of the extant literature.

False memory effects can be viewed as interference effects of various types. The interference either can be supplied externally (by the experimenter) or can occur internally. In the former case, the subject's task is to distinguish between or among sources of overtly-presented information, a process termed source monitoring (Johnson, Hashtroudi, & Lindsay, 1993). Examples of the influence of externally-provided information on remembering can be found in the eyewitness memory paradigm (e.g., Loftus, 1979). In an experiment typical of those in the eyewitness paradigm, Loftus, Miller, and Burns (1978) had subjects viewed a slide sequence depicting a traffic accident. In one of the slides in the sequence, a stop sign was present. Subjects were then asked some questions about the
events depicted in the slides. In passing, one question erroneously referred to the traffic sign as having been a yield sign. Subjects were later asked to identify the sign seen in the original event. Interest in such an experiment lies in the extent to which people will report having seen the item that was not present in the original event but was only suggested to have occurred (e.g., the yield sign). An extensive literature exists revolving around this class of experiments in which subjects are asked to report what happened in an original event after encountering misleading postevent information; the general finding is that subjects have considerable difficulty disregarding the information contained in the subsequent suggestion and report having seen such information in the original slide sequence (see Johnson et al., 1993; Loftus, 1993 for reviews).

Memory confusion errors need not arise from externally-presented information, however. There are several lines of work showing that internally generated information later can be misattributed to having occurred in an external event. For example, Johnson, Raye, Wang, and Taylor (1979) demonstrated that asking subjects to imagine objects enhances the frequency with which they believe they saw the objects. The central question in these types of experiments centers around reality monitoring (Johnson and Raye, 1981). Essentially, the question subjects must answer is: Did I see, hear, or experience the event in question, or did I just think about, imagine, or infer it? Johnson’s research shows that people confuse what they previously imagined (or thought) with what actually occurred externally.

Not all experiments exploring how internal processes can induce false memory involve explicit requests for effortful processing (e.g., imagery) by the subject. In some
cases, investigators have explored more spontaneously-occurring effects, such as how the inferences people draw affect what they remember. For example, Johnson, Bransford, and Solomon (1973) showed that when subjects are presented with a short description containing the clause “He was pounding the nail,” they will often later recognize a sentence containing “He was using the hammer” as being exactly the same sentence as the one heard in the study phase (see also Brewer, 1977; Bransford, Barclay, & Franks, 1972; Harris, 1974).

Within this domain of spontaneously-occurring effects are associative effects in recognition and recall. The experiments reported here fall into this category. Below, I give a brief overview of some of the early experiments in this domain, then review in some detail an experiment reported by Deese (1959b), which forms the basis for the experiments reported here. I then describe recent work by Roediger and McDermott (1995), which extended Deese’s (1959b) findings in several ways. I then provide the motivation for the current experiments, which build upon those of Roediger and McDermott (1995).

**Associative Effects in False Recognition**

Underwood (1965) reported a classic study showing how associates of a word can elicit false recognition of the word. He employed a continuous recognition task: Subjects heard a long series of words presented in succession, and for each word subjects were to decide whether it had occurred previously in the list. The basic finding was that people would claim to recognize a non-presented word (e.g., chair) more often when it had been preceded by an associatively-related word (e.g., table) than when no related words had been earlier presented. In Underwood’s (1965) report, subjects recognized nonpresented
associates of presented words .28 of the time; the false alarm rate in the absence of previously-presented associates was .14; therefore, the difference (.14) is the measure of interest, the false recognition effect.

False recognition in the continuous recognition paradigm has been replicated numerous times (e.g., Anisfeld & Knapp, 1968; Kimble, 1968). Synonyms can produce the effect, and so can category exemplars (e.g., *robin* can elicit later false recognition of *bird*). Antonyms have sometimes been found to lead to false recognition, although the evidence here is mixed (Anisfeld & Knapp, 1968; Underwood, 1965). The false recognition effect is enhanced when several items (e.g., *rest, dream, bed*) converge on a single associate (e.g., *sleep*) relative to the conditions in which a single associate (e.g., *rest*) is presented once or 3 times (Hall & Kozloff, 1973; see too Hintzman, 1988; Shiffrin, Huber, and Marinelli, 1995). The effect has been obtained for laboratory-established associations, as well as associations existing pre-experimentally in semantic memory (Saegert, 1971; Wallace, 1967; but see Hall & Ware, 1968; Wallace, 1969 for failures to obtain the effect). In addition, the false recognition effect has been extended to a standard (study-test) recognition procedure in which subjects are presented with discrete study and test phases: Test items that are associatively related to items presented in a study phase receive more false alarms than test items with no strong associative links to the studied items (Saegert, 1971).

The effects of false recognition obtained from prior presentation of one or two associated words have been usually found to be quite small (e.g., Paul, 1979); indeed, some researchers have failed to obtain the effect at all (e.g., Gillund & Shiffrin, 1984). In
general, with this procedure subjects erroneously recognize about .10 of the control words, which are unrelated to previously-occurring words, and about .18 of the words associatively related to the previously-occurring words (Vogt & Kimble, 1973); the difference in these false alarm rates (.08) is the false recognition effect. Perhaps because the effect of associates on recognition is variable and (when obtained) small, few researchers continued to study it after the mid-1970’s.

Most experiments examining how associations can induce false memories have used recognition (either the continuous recognition procedure or the standard, study-test procedure) as the criterion task. In fact, most constructive memory experiments in general have used false alarms in recognition as the dependent measure. However, as I will explain in the following section, if it were possible to obtain predictable memory errors on a recall test, it would be desirable to do so. Such a finding would allow a more purified examination of false recollection. The basis for this assertion lies in the generally accepted assumption that recognition is based upon two processes: familiarity and recollection (Atkinson & Juola, 1973; Jacoby, 1983a, 1983b; Mandler, 1980; 1983; see Mandler, 1990 for a review). That is, a person may claim to recognize an item on a recognition test either because it “seems familiar” or because the person can recollect some specific aspect of having previously experienced it. It is the breakdown of this second process, producing the illusion of remembering, that is more central to the study of false recollection.

The primary question of interest is whether and how people come to consciously recollect events that never previously occurred; a vague feeling of being familiar with such events is less mysterious. The latter experience can be explained by semantic activation
theories (e.g., Anderson & Bower, 1973; Collins & Loftus, 1975; Collins & Quillian, 1972), which hold that each encounter with a concept (e.g., a word) elicits activation of related concepts (e.g., associated words); such activation leads to a feeling of familiarity with the non-occurring items. The more often a concept is activated (either through explicit contact or implicit activation through relatedness), the more familiar the concept will seem.

Thus, false recognition effects can be easily accounted for by invoking the concept of familiarity. The more interesting question is whether and how people might come to recollect the presentation of the nonpresented associates. Underwood (1965) assumed that his false recognition experiments were in fact studying recollection, or, in more modern terms, failures of reality monitoring (e.g., confusions in whether one heard *table* or just thought about it when presented with *chair*). Underwood (1965) argued that experiencing a word often induces people to think consciously of a related word; this phenomenon he termed the implicit associative response (IAR). Underwood conceived of the IARs as “actually occurring” to the subjects during the study episode (Underwood, 1965, p. 122). The associative response was implicit in the sense that it was not overtly produced; nevertheless, the subject was thought to be fully aware of the production of this item. False recognition was said to arise as a result of a later failure in discriminating between items heard and those thought of (the IARs); Underwood (1965) claimed that subjects become “confused” (p.128) about which was presented and which only thought about. Therefore, false recognition was thought to be a (faulty) recollective experience. However, as evident from dual process theories of recognition as reviewed above, the
possibility exists that Underwood was not measuring inaccurate recollection but was instead studying familiarity. Indeed, Kintsch (1968) argued that recognition measures familiarity only, and not recollection: “recognition does not involve retrieval,” he asserted (p.485). The currently accepted view seems to be that recognition involves some mixture of familiarity and recognition. At any rate, because free recall is thought to be a purer measure of recollection than is recognition (Tulving, 1985), it would be especially interesting if false recall effects could be induced by presentation of associates.

One might wonder why Underwood and others did not attempt to obtain false memory effects in a recall task. This trend is probably attributable to the belief that it is very difficult to obtain errors (especially systematic errors) in free recall unless subjects are specifically instructed to guess (Cofer, 1967; Roediger & Payne, 1985). Intrusion rates in recall of lists of unrelated words or pictures are usually very low. As Cofer (1973) remarked, “subjects in recall experiments seem reluctant...to produce material when they are uncertain that it is correct” (p.538). For this reason, one might expect that it would not be possible to induce false recall of words by presenting associatively related words.

**Associative Effects in False Recall**

There is, however, one generally overlooked report of robust intrusions on a free recall test induced by associations to presented words. Although Deese (1959b) was not interested in errors *per se*, he reported a procedure with which false recall can be elicited through associations. Deese’s (1959b) interest was in using such errors as converging evidence for the importance of the associative structure of to-be-remembered material in determining what a person recalls. He had shown in previous work that the strength of
associations within a list correlated positively with level of accurate recall (Deese, 1959a). That is, if word lists of a given length (e.g., 50 words per list) are constructed by randomly choosing words from the English language, there will be some associative relations between some of the words within each list. In addition, the degree to which the relations exist within lists will vary across different lists of randomly selected words. Deese (1959a) demonstrated that the degree to which items in a list were associatively related to one another correlated with the overall level of recall for the list. He also noted that intrusions sometimes occurred on recall tests and that there were no theories to explain the intrusions. Deese (1959b) proposed that the associative structure of the list could predict not only accurate recall but also intrusions. Therefore, he conducted an experiment to determine whether word association norms could predict the occurrence of intrusions in free recall.

In order to try to find a correlation between intrusions in recall and list structure, Deese (1959b) needed to find a way to obtain substantial levels of intrusion in recall. To this end, he constructed 36 lists, each of which was comprised of 12 associates to a critical nonpresented word. For example, he presented subjects with bed, rest, awake, tired, dream, wake, night, eat, sound, slumber, snore, pillow (i.e., the 12 most commonly-produced words when subjects are asked to free associate to the word sleep). He then measured the probability with which subjects would erroneously recall the critical nonpresented item (e.g., sleep) on an immediate free recall test. Although all his lists were comprised of the 12 most common forward associates to the critical item, the lists differed in their backward association strengths. That is, although the lists were constructed by
collecting the words most frequently produced in response to the critical item on a free association task (i.e., the forward associates), the mean probability with which the list words (e.g., bed, rest) elicited the target (e.g., sleep) as an associate (i.e., the backward association strengths) differed across word lists.

Deese predicted that if the associative structure of the list influenced errors (as well as correct recall), then systematic fluctuations in the associative structure might affect the probability of intrusions in the recall test. Specifically, he predicted that the backward association strength of the lists should correlate with the probability of obtaining intrusions of the critical items on free recall tests given immediately after presentation of each 12-word list. Deese's (1959b) predictions were confirmed; Figure 1 shows that the probability of erroneous recall of the critical nonpresented item, which varied from 0% to 44% across lists, was highly correlated ($r = .87$) with the backward association strengths of the lists, or the mean probability with which the individual list words elicited the target item on a free association task.

Deese (1959b) concluded that the associative structure of a list was critical in determining accurate and false recall. Since publication of his report, however, very few people have pursued his finding. Cramer (1965) did report a similar finding, although it was not the primary focus of her paper. She presented subjects with 26 words: 6 filler words and 4 sets of 5 associatively related words. The associative sets converged upon 4 critical nonpresented words. Cramer (1965) reported that 51% of her subjects erroneously recalled at least one (of the 4 possible) critical words.
Figure 1. The probability of false recall of critical nonpresented associates in Deese’s (1959b) study as a function of the mean probability with which individual list items elicited the associates in a free association task.

Roediger and McDermott’s (1995) approach

Roediger and McDermott (1995) were interested in using associatively related words as a means to study false recall. We employed the lists that do elicit reliable, predictable intrusions to learn more about the phenomenon. Our first goal was to try to replicate Deese’s (1959b) finding. We found it highly surprising that he was able to elicit such robust levels of false recall (at least on some of the lists) given that (1) the materials used were word lists, which are generally not conducive to production of intrusions; (2) the lists were very short; (3) the test was a free recall test, which is usually resistant to intrusive errors; and (4) the test was given immediately after list presentation. Therefore,
we attempted to replicate Deese’s basic finding under conditions in which subjects were instructed explicitly not to guess.

In our first experiment, we replicated Deese's (1959b) surprising finding of robust intrusions on immediate free recall tests. We used the six lists that produced the highest intrusion rates in Deese's experiment and presented them auditorily; our subjects produced the critical intrusions on 40% of the lists. In addition, we measured false recognition on a final yes/no recognition test. Subjects classified the critical nonpresented words as "old" (i.e., studied) 84% of the time; this rate was markedly similar to the hit rate (86%). False alarms to unrelated lures occurred only 2% of the time. Therefore, presentation of multiple words all converging on a single nonpresented associate allowed us to obtain robust false recognition effects (in contrast to the weaker and somewhat inconsistent effects obtained in the procedure in the Underwood tradition described previously). However, the possibility existed that the process underlying this recognition effect was a strong feeling of familiarity, and not false recollection; this point was addressed in a second experiment. Note, however, that the critical items were produced .40 of the time on an immediate free recall test, which is thought to be a relatively direct measure of recollection (Tulving, 1985).

Roediger and McDermott’s (1995) second experiment had three main purposes. First, we wanted to extend the false recall and recognition phenomena to a number of other word lists. Therefore, we constructed 24 lists, all comprised of the top 15 associates to our chosen critical nonpresented words.

Our second objective was to examine the role of initial recall on later recognition
of critical lures. In our first experiment, subjects took immediate free recall tests after each of the 6 study lists and then took a final recognition test. To what extent was the high false alarm rate to critical lures on the recognition test attributable to the prior recall tests (40% of which had included production of the critical item)? To answer this question in Experiment 2 we tested only half of the lists with an immediate free recall test; the other half were followed by math problems for an equivalent time period. All presented lists were later tested on a final recognition test.

The rationale behind Roediger and McDermott's manipulation of prior testing history is crucial to understanding the motivation behind the current experiments. The idea is grounded in the phenomenon of the testing effect: Retrieval of studied items is enhanced by taking a prior test, relative to a condition in which no initial test occurs (Hogan & Kintsch, 1971; Spitzer, 1939; Thompson, Wenger, & Bartling, 1978). That is, if two groups of subjects study a list of words, and one group receives an initial test while the other group performs some other activity, the group receiving the initial test will remember more on a later test than the group not taking an initial test. In fact, a group of subjects taking a test will sometimes later outperform another group that is permitted to re-study the entire set of material (Hogan & Kintsch, 1971). The testing effect demonstrates that the act of retrieval is not simply diagnostic of memory; it can sometimes dramatically influence what one later remembers. This effect of testing is generally beneficial (Wheeler & Roediger, 1992; see Roediger & Guynn, in press, for a partial review), with a prior test enhancing recall on a later test.

Thus far, most of the literature on the testing effect has focused on veridical
memory. Similarly, the literature on false memory remains generally separate from and uninformed by the repeated testing literature (see, however, Roediger, Wheeler & Rajaram, 1993; and Poole & White, 1991, for some exceptions). Because testing can have profound effects upon veridical memory, it seemed likely that testing might also enhance memory for events that did not occur but are associated with events that did occur. Roediger and McDermott (1995) wondered whether a testing effect could be obtained for the critical nonpresented items using the paradigm introduced by Deese (1959b). Such an effect would inform our understanding of the phenomenon of illusory recollection. If the critical nonpresented items generally behave as if they had been presented in the study list, it would be expected that a testing effect would occur for these items (i.e., false recall on a first test would enhance false recognition on a later test, relative to a previously nontested condition). In addition, to the extent that results from this paradigm generalize to accurate and erroneous retrieval of everyday events outside the lab, such a finding would have important practical implications. Therefore, Roediger and McDermott (1995, Experiment 2) asked whether taking an initial test might facilitate later recognition of both studied items and critical nonpresented items.

The third objective in Roediger and McDermott’s (1995) Experiment 2 was to address the familiarity/recollection issue alluded to previously. To do this, we decided to explore the phenomenological experience of subjects when recognizing the critical nonpresented items. Tulving (1985) proposed that people have two means of accessing their personal past: remembering and knowing. Remembering an event is the phenomenological experience of being able to mentally re-experience the original event.
When we are sure that an event happened, but we lack the phenomenological experience of being able to re-live the details and context, we are said to know that it happened. Tulving (1985) proposed that this distinction could usefully be applied to laboratory paradigms in order to gain further understanding of subjects' phenomenological experiences during retrieval, and many researchers have done so successfully (e.g., Gardiner, 1988; Rajaram, 1993; see Gardiner & Java, 1993 for a review). When used in a recognition memory paradigm, the remember/know procedure occurs as follows. Subjects are told that for each word they see, they should decide whether the word is old (i.e., was studied) or new (i.e., was not studied). They are further instructed that for every item they judge to be old, they should make an additional classification: remember or know. They are told that remember judgments should be reserved for those items for which they can recollect some specific aspect of the occurrence (e.g., something distinctive in the speaker's voice as he or she said the word, the word preceding it, or other features); when subjects lack this experience, they are instructed to respond with a know judgment.

The goal in using this procedure was to purify the recognition judgments by examining remember responses. The critical question was whether subjects would ever claim to recollect the instance of presentation of nonpresented items (by assigning remember judgments to them). If subjects were simply responding on the basis of a general, vague feeling of familiarity, we would expect them to classify most such false alarms as know judgments. Indeed, the existing literature shows that subjects almost always classify false alarms as being known (Gardiner & Java, 1993). Of course, this makes sense in that in the case of false alarms there is no original episode to remember; if
an item is erroneously recognized, it is likely due to a vague feeling of familiarity. One purpose of Experiment 2 was to discern if, with related words, a different pattern would be obtained such that subjects would mistakenly claim to remember the presentation of the critical nonpresented words.

To answer these questions, Roediger and McDermott (1995, Experiment 2) presented subjects with 16 lists; after each of 8 lists, subjects took immediate free recall tests, and after each of the other 8 they worked math problems. After all 16 lists had been presented, subjects took a final recognition test during which they classified each word as old or new; in addition, remember/know judgments were collected for all words classified as old.

On the immediate free recall tests, subjects produced the critical nonstudied item 55% of the time. Thus, we were able to devise our own lists that worked as well as (actually, better than) those used by Deese (1959b) in producing false recall. The second interesting result is that the act of recall enhanced later recognition of items. As shown in the top third of Table 1, recognition of studied items from the lists that had been recalled revealed higher hit rates than those that had been followed by math problems (.79 and .65, respectively). The data in the middle third of the table demonstrate that the same pattern was found for false alarms to the critical nonpresented items (.81 and .72). These data indicate that the act of retrieval seems to augment later false recollections in much the same way that it enhances veridical retrieval (see also Roediger, et al., 1993). Therefore, taking a test enhances the later probability with which people will claim to be able to vividly recollect the original presentation episode for both studied and critical nonstudied
items. This finding is true despite the emphasis in the remember/know instructions that the judgments should be made with respect to the study episode (i.e., not whether subjects remembered or knew that they had produced the item on the immediate free recall test).

Table 1. Recognition results for studied items (top), critical lures (middle), and standard lures (bottom) in Roediger and McDermott's (1995) Experiment 2.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Recall Test?</th>
<th>Proportion of items</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Old</td>
<td>Remember</td>
<td>Know</td>
</tr>
<tr>
<td>Studied</td>
<td>Yes</td>
<td>.79</td>
<td>.57</td>
<td>.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.65</td>
<td>.41</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>Critical Lure</td>
<td>Yes</td>
<td>.81</td>
<td>.58</td>
<td>.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.72</td>
<td>.38</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>Standard Lures</td>
<td>Yes</td>
<td>.11</td>
<td>.02</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.16</td>
<td>.03</td>
<td>.13</td>
<td></td>
</tr>
</tbody>
</table>

Surprisingly, we found that subjects often classified the erroneously recognized critical items as being remembered. Furthermore, they made this classification for the critical items as often as they did for items that actually had been present in the list.

Subjects were not simply classifying most items as being remembered, however; they almost always assigned know judgments to standard, unrelated lures when they were erroneously recognized.
Attempts to replicate the testing effects found by Roediger and McDermott (1995) by using the same paradigm, with a few modifications, have produced variable results. In an experiment that conformed extremely closely to Roediger and McDermott's (1995) Experiment 2, Payne, Elie, Blackwell, and Neuschatz (1996) failed to obtain a testing effect for the critical items, although they did observe a benefit from prior testing of the studied items. One of the goals of the present experiments was to attempt to confirm the existence of the testing effect for critical nonpresented items.

In summary, the powerful false memory results obtained by Roediger and McDermott (1995) demonstrate that this paradigm a useful one with which to study the development of false memories. The paradigm complements other methods of studying false memory and offers several advantages over some commonly-used procedures. For example, it allows one to obtain reliable, predictable intrusions on a free recall test, which, for reasons discussed previously, is preferable to obtaining false recognition. In addition, the paradigm allows one to study spontaneously-occurring processes, with no interfering events supplied by the experimenter. That is, it allows us to study the phenomenon of naturally-occurring interference from semantic memory to episodic memory. In addition, with this procedure, it is possible to obtain many more observations per subject than in some false memory studies (e.g., the misinformation paradigm, in which it is often possible only to obtain one observation per subject); this feature has the obvious advantages of adding power to experiments and allowing variables to be manipulated within-subjects. Finally, the use of word lists allows for an application of our considerable knowledge base from the single trial free recall word list paradigm to the study of false memories. The
benefit of this feature will become apparent in the explanations of the rationale for the current experiments.

**The Goals of the Present Experiments**

The experiments reported here use Roediger and McDermott's (1995) paradigm to explore further the phenomenon of erroneous recollection, with special attention given to the importance of an initial test on later retrieval of nonpresented items.

Because the paradigm used here is relatively new, the aim is to investigate its parameters through a series of experiments rather than to test specific theories about the false memory effects. Some of the questions addressed include: What is the effect of delaying recall on later memory for the critical nonpresented items? That is, is there “forgetting” of nonpresented items as there is for presented items? Second, does the repeated recalling of a list increase the likelihood of retrieval of the critical nonpresented items? Does repeated retrieval increase remembering of nonpresented items? Third, if subjects are instructed to guess during an initial test but then asked to report only the actual study episode (without guessing) on a later test, are they able to distinguish previous guesses from authentic memories? Finally, with multiple study-test presentations, are subjects able to edit out the critical intrusion after several recall trials? The rationale for asking each of these questions is provided in the introduction to each of the experiments.

**Experiment 1**

In Experiment 1 the effects of retention interval on the false recall of non-presented associates was examined. Both a short (30 s), filled delay and a longer, 2-day
delay were employed. The logic behind both conditions is explained in the following section, with the short delay considered first.

Roediger and McDermott (1995, Experiment 2) found that on an immediate free recall test, subjects produced the critical lure 55% of the time. Items that had occurred on the study list were produced 62% of the time. Figure 2 shows the probability of recall of studied items plotted as a function of position of occurrence in the list. There were large primacy and recency effects, replicating prior work (e.g., Murdock, 1962). A comparison of the probability of recall of the critical nonpresented items (depicted in the figure as the dashed line) and the probability of recall of items in the middle portion of the list shows comparable levels of recall in the two conditions. This finding is interesting in that recall of items from the middle of the list is usually thought to represent retrieval from long term store (Atkinson & Shiffrin, 1968) or secondary memory (Waugh & Norman, 1965). Thus, we can say that the critical nonpresented items are recalled from long term store with the same probability as items actually occurring in (the middle portion of) the list.

Postman and Phillips (1965) and Glanzer and Cunitz (1966) showed that the introduction of a short, filled delay leads to attenuation of recall (forgetting) relative to an immediate test. The effect is localized in the last portion of the serial position curve; that is, such a delay eliminates the recency effect but does not affect recall of the rest of the list. Because the recency effect is thought to represent retrieval from primary memory, we can say that introduction of a short, filled delay affects recall from short term store, leaving recall from long term store relatively unharmed. On the basis of the assumption that the critical nonpresented items are recalled from long term store (because they are
recalled with the same probability as if they actually had been presented in the middle of the list), we can predict that a short, filled delay will have no effect on recall of the critical items. The comparison of performance on an immediate test and a test followed by a short, filled delay was intended to test this hypothesis.

Figure 2. Probability of recall as a function of serial position in Roediger and McDermott's (1995) Experiment 2. (The dashed line indicates the mean probability of recall of the critical nonpresented items).

For reasons explained below, a longer (2 day) delay was also introduced in this experiment. Therefore, after studying 24 lists (8 of which were followed by an immediate free recall test, 8 by a delayed free recall test, and 8 by no initial test), subjects returned 2 days later and were asked to recall all the words they could retrieve from the presented lists. There were two questions addressed by this test: (1) What would happen to false recall, relative to veridical recall, after a substantial delay; and (2) Would the answer to the first question differ as a function of the test condition on the first day? That is, would
taking a first test (either immediately after list presentation or after a short delay) affect performance after a 2 day delay relative to a condition in which no initial test was taken?

It is generally assumed that introduction of a delay between study and test diminishes the probability of recall of an individual item. However, there is some evidence in the prose memory literature to suggest that false recall might actually increase over long delays, albeit in paradigms quite different from the one used here. Spiro (1980) found that false recall of prose passages increased across a 6-week period. In addition, Sulin and Dooling (1974) found false recognition of schema-consistent prose passages to increase over a delay of one week; Barclay and Wellman (1986) extended this outcome to one year. Most relevant to the current experiments is the work of Leicht (1968), who presented subjects with several associates and then examined intrusions on free recall immediately and after 24 hours; there was forgetting for the presented words and either no change or possibly an increase in intrusions across the delay. (These conditions and this comparison were not central to the article; therefore, the relevant statistics were not reported.)

The 2-day delay condition in the present experiment also allowed for an assessment of whether any differences found between the initial test conditions (an immediate test or a test given after a short, filled delay) might be carried over onto a final free recall test given much later. In addition, the final free recall test permitted an assessment of whether the act of recall on Day 1 would influence recall levels 2 days later. In other words, would recall of items from the lists that had been tested on Day 1 differ from recall of those not initially tested? As reviewed previously, evidence from the
literature on the testing effect suggests a positive effect of testing would occur for studied items. The interesting question was whether such an effect might also occur in recall of critical items. Roediger and McDermott (1995) obtained a testing effect for critical lures on a recognition test given immediately after study of all the lists, but Payne et al. (1996) and Schacter, Verfaellie, and Pradere (1996) failed to replicate this outcome. The question here was whether a testing effect would extend to a final free recall test given after a 2-day delay.

In summary, the objectives of Experiment 1 were to determine (1) whether (and how) introduction of a short, filled delay would affect recall of the critical nonpresented items; (2) how a substantially longer, 2-day delay would affect false recall in relation to accurate recall; and (3) whether the answer to this second question would differ as a function of whether an initial test had occurred in the first session (and whether the timing of the initial test--immediate or 30-s delayed--would influence later false recall).

Method

Subjects

Forty-five Rice University undergraduates and summer students participated for pay, at $5 per hour.

Design

The nature of the initial test was manipulated in 3 within-subjects conditions (Immediate, Delayed, and No Test). Twenty-four study lists were divided arbitrarily (with the stipulation that intrusion rates obtained in previous experiments were approximately equivalent across sets) into 3 sets of 8 lists for purposes of counterbalancing. Across
subjects, each set (and therefore each list) was assigned to each condition equally often.

**Materials**

Twenty-four lists of 15 items each were used for this experiment; they were the same lists as those reported in the Appendix of Roediger and McDermott's (1995) article. There was no overlap in words across lists, and no critical items appeared in any list. Ordering of words within lists was held constant, and the words most strongly associated with the critical items generally occurred toward the beginning of the lists.

Lists were presented on a tape recorder in a male voice at a rate of approximately 1.5 sec. Because all subjects heard all lists, only one tape was necessary. Initial test condition occurred in a mixed fashion.

Subjects were provided with 2 stacks of 4" x 11" sheets of paper; one stack contained lined sheets for recall, and one contained sheets with math problems.

**Procedure**

Subjects were tested individually or in groups of 4 or fewer. They were seated at tables with the two stacks of paper in front of them. They were told that the purpose of the experiment was to examine the relation between their memory and math abilities. It was explained that they would hear a series of lists (24 in all) presented via the tape player and that they should pay close attention to the lists because after some lists their memory would be tested.

Following presentation of each list, subjects either (1) took an immediate test (i.e., spent 90 s recalling the immediately-preceding list and then did 30 s of math problems), (2) took a delayed test (i.e., did math problems for 30 s and then were asked to recall for
90 s), or (3) took no test (i.e., did math problems for 30 s and then were asked to do an additional 90 s of math problems). There were always 2 min between list presentations, and subjects never knew until after a list had been presented whether they would first recall it or do math problems. Furthermore, when a list was followed initially by math problems, subjects did not know until the end of the math period whether they would then be asked to do more math problems or to recall the list.

Instructions given to subjects were designed to disguise the fact that the math problems were included only to instantiate three test conditions (Immediate, Delayed, and None) in hopes that subjects would take the math problems seriously (and thereby not rehearse the list in the delayed condition) and that they would not realize that a final test would be given when they returned 2 days later.

Subjects were informed that they would perform 2 tasks after presentation of each list. They were told that immediately after list presentation, the experimenter would say "recall" or "math" to indicate what should be done. When the experimenter said "recall," subjects were to take one of the lined sheets in front of them and write down as many words as they could recall from the immediately-preceding list. The experimenter stressed that although the object was to remember as many words as possible, subjects should be confident that everything they wrote down had occurred in the list. They were told that they would be given 1.5 min in which to recall the list, and the stopwatch would beep when this time had passed. When the stopwatch beeped, they were to turn the sheet over. Subjects were further instructed that when the experimenter said "math," they should take the top sheet from the math stack and work problems as quickly and accurately as possible.
until the stopwatch beeped, at which time they were to turn the sheet over. They were informed that sometimes they would recall and then do math, or do math and then recall; furthermore, they were told that sometimes they would be asked to do math problems twice in a row or to recall twice (this latter condition never actually occurred). It was stressed to the subjects that the experimenter would always tell them what they should be doing and when to do it. Questions were solicited and answered.

Upon completion of the last list and its corresponding test and/or math periods, subjects were dismissed for the day and reminded to return 2 days later for a further, unspecified experiment.

When they returned, subjects were given lined sheets of paper (with 200 lines per subject) and asked to recall all the words they could retrieve from the first session. They were asked to recall as many words as possible, regardless of whether the list (or the individual word) had been recalled on the first day. As in the first session, they were asked to be certain that every word they wrote down had occurred on the tape player. They were given 15 min to perform the task; after each min, a tone sounded on a computer, signaling them to draw a line under the last item recalled and to continue attempting to recall until the session was over.

Results and Discussion

Initial Tests

Mean proportions of studied and critical nonstudied items recalled on the initial tests are shown in the left half of Table 2. Several observations can be made from examining the table. First, although a short, filled delay attenuated the level of veridical
recall (compared to the immediate test condition), the delay had no effect on the recall level of the critical items. For studied items, the difference in recall proportions in the immediate condition (.58) and delayed condition (.50) was reliable, $t(44) = 7.71$, $SEM = .009$. As shown in the serial position curve of Figure 3, this difference was localized in the last few items presented. That is, the 30 s filled delay eliminated the recency effect for studied items, replicating prior research (Glanzer & Cunitz, 1966; Postman & Phillips, 1965).

Table 2. Proportions of studied items and critical nonstudied items recalled on the initial tests and the final free recall test as a function of initial test condition in Experiment 1.

<table>
<thead>
<tr>
<th>Initial Test Condition</th>
<th>Initial Tests</th>
<th>Final Free Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studied</td>
<td>Critical NS</td>
</tr>
<tr>
<td>Immediate</td>
<td>.58</td>
<td>.44</td>
</tr>
<tr>
<td>Delayed</td>
<td>.50</td>
<td>.46</td>
</tr>
<tr>
<td>No Test</td>
<td>----</td>
<td>-----</td>
</tr>
</tbody>
</table>

Although robust levels of false recall were obtained in both conditions, there was no reliable difference between the conditions; the critical items were produced 44% of the time on the immediate test and 46% on the delayed test, $t(44) < 1$. The mean proportion of critical items recalled (collapsing across conditions) is represented in Figure 3 as the
dotted line. (Note that the serial position designation is meaningless for these items because they were never presented.) As apparent from an examination of Figure 3, subjects recalled the critical nonstudied words with a probability comparable to recall of items presented in the middle portion of the list, which is usually thought to represent recall from long term memory (e.g., Atkinson & Shiffrin, 1968; Glanzer & Cunitz, 1966).

Figure 3. Probability of recall of studied items on the initial recall tests (both immediate and delayed) in Experiment 1 as a function of serial position of the presented items. (The dashed line indicates the mean probability of recall of the critical nonstudied items [collapsed across test condition].)

There was a significant interaction between the test conditions and item type, $F(1,44) = 5.63$, MSE = .02. As discussed above, delaying the initial test affected production levels of list items but not critical intrusions. Another way of looking at the interaction is that on the immediate test, accurate recall exceeded false recall, $t(44) = 3.26$, SEM = .04, but after a short delay, the probabilities of false and accurate recall were not
significantly different, t(44) = 1.07, SEM = .045.

Roediger and McDermott (1995) reported that on an immediate test, when the critical items were produced, they tended to occur toward the end of the subjects’ output for the list: 63% of the time they occurred in the last fifth of subjects’ output. An identical analysis for the immediate condition in the present experiment showed similar, but less dramatic results: 38% of the critical items occurred in the last fifth of subjects’ recall protocols. For the lists on which the critical item was recalled, the mean recall position was 6.2 (out of 9.6 items written down), or 65% of the way through their output. Figure 4 shows cumulative proportions recalled as a function of output quintiles for the immediate and delayed conditions of this experiment. The basic finding emerging from these data is that when the critical nonpresented item is produced on an immediate free recall test, it tends to appear toward the end of subjects’ output.

An examination of the output position data for the 30-sec delayed test shows a different pattern of results. When the cumulative proportion recalled is plotted against output quintile, the resulting function is almost perfectly linear, as shown in Figure 4. This finding indicates that critical items did not occur disproportionately in any particular section of the output protocols. The mean output position was 4.9 (out of 8.6 items produced on these tests), or 57% of the way through the subjects’ output.

The discrepancy in output positions of immediate and delayed tests makes sense if one assumes that a major factor contributing to the late occurrence of critical intrusions found on the immediate test is that subjects begin their recall by producing the items in primary memory, those presented at the end of the list. This recall strategy allows
Figure 4. Cumulative probability of recall of the critical nonpresented item as a function of output position in recall for the immediate and delayed tests of Experiment 1. Quintiles refer to the first 20% of the subject's responses, the second 20%, etc.

little chance for the critical item to appear in the first part of the recall protocol. When a 30-s filled delay is introduced, and primary memory is removed from subjects' recall, more freedom is permitted in the output position of the critical nonpresented item.

An analysis of subjects' errors showed that the critical nonpresented items were not the only items falsely recalled by subjects. However, of all possible words in the English language, subjects intruded an average of only .22 items per list on the immediate tests and .32 items per list on the delayed test. That is, subjects produced an item that was neither a studied word nor the critical nonpresented word on about every fifth list on the immediate tests and on every third list on the delayed test. This difference was reliable, \( t(44) = 2.8, SEM = .036 \). In general, the intrusions were semantic associates of the presented words. Occasionally, however, subjects wrote a word phonetically similar to a
presented item.

In summary, results on the initial tests showed robust levels of false recall, replicating Roediger and McDermott's (1995) finding. Furthermore, there was no effect of a short, filled delay on levels of false recall. However, the standard finding that such a delay attenuates accurate recall by eliminating the recency effect was obtained. An analysis of where the critical items occurred when they were produced showed that they generally occurred toward the end of the output for the immediate recall condition (replicating Roediger & McDermott, 1995) but appeared equally distributed throughout the recall protocols for the delayed condition. Finally, intrusions other than the critical nonpresented items occurred infrequently, indicating that subjects were not simply guessing or free associating during the recall period, thereby following their instructions not to guess.

Final Test

Results from the final free recall test (given 2 days later) are reported on the right side of Table 2. The overall proportions of studied items and critical items recalled on this test were lower than on the initial tests. However, caution should be used when making this comparison because the tests differed not only in delay but also in that in the first session, an initial test was given after presentation of each of the lists, whereas two days later, the test was a final free recall test covering all previously-presented lists. Therefore, no direct comparisons are made between the data for the two test sessions.

Perhaps the most remarkable aspect of the final free recall data is that the overall proportion of critical items recalled ($M = .20$) exceeded the proportion of studied items
recalled ($M = .12$). A 3 (Initial Test) x 2 (Item type) within-subjects ANOVA showed this main effect to be significant, $F(1, 44) = 21.28$, $MSE = .02$. The test condition of Day 1 also exerted a significant effect on later recall, $F(2, 88) = 42.56$, $MSE = .01$. There was no reliable interaction between test condition and item type, $F(2, 88) < 1$, indicating that studied items and critical items behaved similarly as a function of prior testing history:

Regardless of whether the lists had been tested immediately after presentation, after a 30 s delay, or not at all, the proportion of list items recalled 2 days later was lower than the proportion of critical intrusions.

Planned comparisons showed that for studied items, subjects recalled more items when they had previously taken an immediate test than when the initial test had been delayed, $t(44) = 2.06$, $SEM = .012$. However, for critical items, there was no difference in recall between the immediate and delayed conditions, $t(44) < 1$. This lack of a difference is not too surprising in light of the fact that the immediate/delay manipulation had no effect on false recall on the initial test. In addition, the immediate and delayed conditions combined produced greater recall probabilities later than did the no test condition for both studied items, $t(44) = 10.59$, $SEM = .01$ and critical items, $t(44) = 5.18$, $SEM = .02$. Thus, the standard benefit from prior testing—the testing effect—was obtained for studied items; more interestingly, the same effect occurred for critical nonpresented items.

It is informative to examine the serial position functions in final free recall as a function of the three initial test conditions, as shown in Figure 5. Because overall recall levels were low (and therefore noisy across serial positions), this curve has been smoothed by averaging 3 successive points to obtain each point (except the first and the last). For
Figure 5. Probability of recall of studied items on the final free recall test in Experiment 1 as a function of serial position and initial test condition. (Dashed lines indicate probability of recall of the critical nonstudied items as a function of initial test condition.)

instance, the mean of the 2nd, 3rd, and 4th positions is represented in the figure as the 3rd position. (The first and last positions in the graph, however, are based on raw data.) It is clear from this figure that false recall proportions exceeded veridical recall proportions in all conditions for almost all serial positions. Recall that Figure 3 showed that the probability of false recall approximated the probability of recall of items that had been presented in the middle portion of the lists. Figure 5 shows that by 2 days later, on a final free recall test, critical items no longer behaved as if they had been presented in the middle portion of the list; instead, they were recalled at levels comparable to the primacy portion of the curve. It is unclear from this experiment how much of this change in relative recall levels is attributable to the test delay and how much to the difference in the tests (final free recall covering all the lists, instead of recall of each individual list). Nevertheless, recall of
the critical nonpresented items differs from recall of items in the middle of the lists.

Depicted in Figure 6 are the cumulative recall curves for both studied items and critical nonpresented items. This graph shows the cumulative probability of recalling the items plotted for the end of each minute of the final free recall test. The first thing to note from this figure is that the critical items were recalled throughout the recall period, not just toward the end. In addition, the proportion of critical nonpresented items recalled is increasing more rapidly than the proportion of studied items recalled, $F(14, 616) = 13.32$, $MSE = .01$. Because one could argue that this interaction might simply be caused by the main effect of item type combined with a floor effect in the first couple minutes of the recall period, an analysis of the last 8 minutes of the recall period was performed. The item type x minute interaction for this section of the curve was also reliable, $F(7, 308) = 4.70$, $MSE = .00$, providing further evidence that the recall curve for critical items is diverging from that for studied items. A final observation derived from this figure is that recall of neither type of item had reached asymptote by the end of the 15-minute recall period, indicating that if subjects had been allowed more recall time, they would have recalled more of both types of items.

Finally, an examination of noncritical extralist intrusions revealed that on average, subjects produced 3.5 such items on their final free recall test. Therefore, after studying 24 lists, 2 days later they correctly recalled an average of 44.4 (of 360 possible) words, produced 4.7 (of 24 possible) critical intrusions, and produced 3.5 other intrusions (out of all other English words).

In sum, results on the final free recall test of Experiment 1 demonstrate the
Figure 6. Cumulative probability of recall of critical nonpresented items and studied items recalled for each minute of the final free recall test in Experiment 1.

robustness of false recall in this paradigm. Indeed, after two days, the proportion of critical nonpresented items recalled exceeded the proportion of studied items recalled. In addition, the experiment confirms the critical role of testing in determining later false recall.

Experiment 2

The results of the first experiment demonstrate the critical role that testing can play in the creation and maintenance of false memories. Experiment 2 was designed in part to further examine false memories with respect to the testing effect, and to explore a related phenomenon: hypermnesia. In addition, the effects that guessing on an initial test can exert on a later test were examined.

The Testing Effect and Hypermnesia

Thus far, discussion of repeated retrieval attempts has been confined to the testing effect: the finding that people generally recall more on a test if they have previously
attempted to retrieve the study material than if they have not. Experiment 2 extended the investigation into the importance of testing by employing a repeated testing paradigm used to study hypermnesia, the improvement in recall over repeated tests (Erdelyi & Becker, 1974; Roediger & Challis, 1989). In this procedure, subjects study pictures or words and are given several successive tests in which they are told, for each test, to recall as many of the studied items as possible. At the level of individual items, some items recalled on (say) the first test will not be recalled on the second test; therefore, forgetting occurs. In addition, some items not recalled on the first test will be recalled on a second test; this phenomenon of item recovery is called reminiscence. When the overall level of reminiscence outweighs forgetting, there is a net increase in the number of items correctly recalled across the successive tests. This net increase is called hypermnesia. Hypermnesia is a fragile phenomenon, with narrow boundary conditions; it generally occurs with long study lists composed of pictures or concrete words processed semantically (Erdelyi & Becker, 1974; Erdelyi, Finkelstein, Herrell, Miller, & Thomas, 1976). Because the associative lists used in the present procedure were short and not restricted to concrete objects, it seemed likely that hypermnesia would not occur for studied items in this procedure. The interesting question, however, was whether an analogous pattern of "recovery" of items not previously recalled might occur for the critical nonpresented items across successive tests. That is, if given 15 associates related to the concept of (say) sleep, followed by 3 successive recall tests (in the absence of an intervening study episode), would subjects be more likely to recall sleep on the third test than on the first?

Part of the focus of this experiment was on the condition in which 3 initial tests
were taken: Would there be changes across the tests? The number of initial tests was also
manipulated (0, 1, or 3) so that performance on a final free recall test could be examined
as a function of number of initial tests to examine the possible power of testing in
enhancing veridical and false recall. The comparison between the 1 test and no test
conditions mirrors the conditions of Experiment 1, except the final free recall test in this
experiment occurred on the same day (rather than 2 days later). The 3-test condition
allowed for an examination of whether this effect could be extended by increasing the
number of initial tests. If testing plays a large role in the production of false memories, we
would expect that the more tests taken, the larger the effect.

Therefore, in Experiment 2, subjects heard 18 lists; after 6 lists they took 3 initial
recall tests, after 6 lists, they took only one test, and after 6 lists they took no initial test at
all. An examination of the condition in which there were 3 initial tests allowed for
measure of a hypermnnesia-like effect for critical items. Following presentation of all lists,
subjects took a final free recall test. This test allowed for a confirmation of a testing
effect, such as obtained in Experiment 1. In addition, it could be determined whether the
number of initial tests further affected false recall (i.e., whether 3 initial tests produced
higher levels of false recall later than 1 initial test).

The Effects of Guessing on Recall

One commonly-expressed concern about findings of an increased production of
studied words across successive free recall tests (hypermnnesia) is the possibility that this
phenomenon is simply an artifact of an increasingly relaxed criterion on the recall tests
(Dywan & Bowers, 1983; Klatzky & Erdelyi, 1985). Such concern led Cofer (1967) to
introduce a procedure he labeled forced recall. In this procedure, subjects are presented with a study list, which is followed by a free recall test. After they have been given a reasonable amount of time for recall, subjects are told how many words were in the study list; they are asked to compare the number of responses they have provided to the number of words presented and to make up the difference by guessing. Because subjects are required (or forced) to match the number of words recalled with the number studied the procedure is called forced recall.

Cofer (1967) found that when subjects were given lists of associated words and were required to generate guesses on forced recall, production of correct responses did not increase compared to correct responses provided by subjects in a comparable free recall condition. This general finding has since been replicated (e.g., Roediger & Payne, 1985). Thus, use of this procedure has led to the conclusion that increases in recall levels across tests are not attributable to criterion shifts. Indeed, Cofer (1967) concluded that "free recall largely, if not entirely, exhausts the 'storage'" of memory (p. 197).

Roediger, et al. (1993) reported a series of experiments originally intended to explore the effects of recall criterion, but the most interesting results from those experiments lie in the area of false memory. In selected conditions of one experiment, subjects viewed a series of 60 pictures presented in the context of a story. Immediately after presentation of the pictures, some subjects took a free recall test and others a forced recall test. After taking the test, subjects were required to assign confidence ratings to each of their responses with respect to the belief that the response had occurred in the original study list. Although level of correct recall was equivalent across free and forced
recall groups, subjects in the forced (but not free) recall condition made two types of errors: (a) they failed to recognize some of their correct responses as being correct and (b) they "recognized" some of their incorrect responses as having occurred during the study phase. Thus, it seems as though the act of guessing influences memory immediately thereafter. In a sense, when asked to guess, subjects supply their own interfering misinformation (Roediger, et al., 1993).

Experiment 2 was designed to extend this finding and apply it to the paradigm used here. Half of the subjects were given standard, free recall instructions for the initial tests, and half were instructed to produce plausible guesses. At the end of the session, on a final free recall test, all subjects were instructed to recall only items they were sure had occurred in the original study sequence. It seemed plausible to predict that when forced to guess on the lists of associates, subjects would produce the critical nonpresented items. Although this in itself is not too interesting or surprising, the goal was to determine whether this act of producing the critical items as guesses on initial tests would enhance the probability with which these items would be recalled on a later test when subjects were instructed to produce only studied items. This outcome would be consistent with the source monitoring literature, which shows that people often remember what happened but misattribute the source of their knowledge (Johnson & Raye, 1981; Johnson, et al., 1993). In the current experiment, subjects could remember that the critical item occurred (it will often have been produced on the forced recall test) but misidentify the source of this memory as being the original study list. Furthermore, the enhancement attributable to guessing was predicted to differ as a function of the number of prior recall tests.
Therefore, it seemed plausible that the effect of guessing would be magnified for the 3 initial test condition relative to the 1 test condition.

A final purpose of Experiment 2 was to pursue Roediger and McDermott’s (1995) finding that the act of recall enhances later probabilities of *remember* responses for the critical nonstudied items. It was argued in the Introduction that the most interesting aspect of false memories is when subjects erroneously recollect specific features of events that never occurred. For this reason, it was argued, researchers should use recall as the dependent measure (or else collect and examine *remember* judgments when using recognition). It has generally been assumed that performance on free recall tests is based almost entirely on conscious recollection, or *remember* experiences (Tulving, 1985). However, since Tulving’s (1985) initial report on the *remember/know* distinction, there have been no published reports of whether this assumption holds true. Tulving’s (1985) data support the claim; he obtained an overall free recall rate of .51, which was comprised of .45 *remember* judgments and .06 *know* judgments; therefore 88% of the items produced on the recall test were labeled as *remembered*. However, the extent to which this result generalizes to other situations is still in question.

Because the interest here is in identifying and studying cases in which erroneous conscious recollection occurs, I employed the *remember/know* procedure on the final free recall test. Subjects were instructed to write an R or K (for *remember* or *know*) next to each item produced on the final free recall test. Subjects were told to make these judgments with respect to the original study episode (i.e., not whether they *remembered* producing the item on an initial test). On the basis of Tulving’s (1985) data, it was
expected that virtually all recalled items would be labeled as *remembered*; furthermore, it was predicted that the effect of the number of prior tests would be mirrored in the *remember* judgments. That is, like overall recall, *remember* judgments would exhibit an increased hit rate and an increased veridical recall probability and an increased false recall probability as a function of number of prior tests.

In summary, I predicted that repeated attempts to recall a list would result in increasing rates of intrusions of critical items on the initial tests, and these intrusions on the initial tests would be carried over into later claims of *remembering* the specific aspects of the presentation episode of these (nonpresented) items. Encouraging subjects to guess on initial tests was expected to enhance these effects. The predicted results would support the general principle that repeated recall enhances false memories.

**Method**

**Subjects**

Sixty-four Rice University subjects participated for course credit or pay.

**Design**

The experiment consisted of a 2 (instructional set of initial tests: forced or free) x 4 (list condition: not presented, presented but not tested, presented and tested once, presented and tested three times) mixed design, with instructional set serving as a between-subjects factor and list condition a within-subjects factor.

**Materials**

The twenty-four lists used in Experiment 1 were used for this experiment. For counterbalancing purposes, the lists were divided arbitrarily into four sets of 6 lists each.
For each subject, one set of 6 lists served in each of the following 4 conditions: presented once and tested three times, presented and tested once, presented but not initially tested, and not presented. Each set served in each condition an equal number of times across subjects. One master ordering of lists was devised, with the stipulations that: (a) no more than 2 lists from the same set (and therefore the same condition) occurred in immediate succession and (b) occurrence of lists from the four sets was distributed evenly across the beginning, middle, and end of the master list. Subjects were provided with 8 1/2" x 11" sheets of paper, each with 15 lines on which recall responses were recorded. In addition, one numbered sheet of paper was provided for responses in the digit monitoring task (described below).

Lists were recorded digitally in stereo in a female voice using a 22050 HZ sampling rate and 16 bit resolution. The average presentation time for each list of 15 words was 15.6 s (ranging from 14.5 to 17.5 s). The lists were presented via LabTec CS 1400 stereo speakers.

**Procedure**

Subjects were tested in groups of 5 or fewer, and all subjects within an experimental session were assigned to the same condition. Subjects were told that they would hear 18 short lists of words presented via a computer (equipped with speakers) and that their job was to pay close attention to the lists because their memory for the lists would sometimes be tested. They were told that after some lists there would be no memory test but that following most lists they would be given 2 min during which they should write down all the words that they could recall from the immediately preceding list.
Subjects were informed that regardless of whether an individual list was tested, they would receive a digit monitoring task after each list. (The purpose of the digit monitoring task was to eliminate the recency effect, thereby allowing the possibility of hypermnnesia across lists; Tulving, 1967, showed that hypermnnesia is not obtained with short lists in the absence of a distractor activity.) Therefore, following presentation of each word list, subjects heard the recorded command “Count the even digits” played on the speakers. Subjects then heard a set of 30 digits presented rapidly (in about 21 s/set), and subjects were instructed to count the number of even digits presented. When the digit presentation was over, they heard “Write down the number of digits,” which was their signal to write the number of even digits they had counted. They then either took a short break and then heard the next list (in the no test condition), or they heard the word "recall," signaling them to recall the last list they had heard. Subjects did not know until the end of a list and its corresponding digit monitoring task whether the list would be tested.

Subjects were given periodic 15 s breaks in which to relax and prepare for the next list. The breaks occurred after the last recall test for each list (or immediately after the digit monitoring task in the no test condition). Breaks were announced by the command “Take a short break” and ended with “Get ready for the next list.”

In the free recall group, the experimenter stressed to subjects that they should recall as many words as possible, but they should not guess (i.e., they should be reasonably confident that everything they wrote down had in fact occurred on the list they just heard). Subjects were given 2 min in which to recall; after 1.5 min they heard a tone
on the computer, which signaled that they had only 30 s remaining. They were instructed to draw a line under the last word they had produced when they heard the tone so that the experimenter could track their recall in the first 3/4 and last quarter of the total recall time. They were asked to continue to try to recall words for the entire allotted time (2 min); that is, they should not stop trying to recall items from the list until they heard the prompt “stop recalling and turn to the next page in your booklet” on the computer speakers.

Subjects in the forced recall group were instructed to spend the first 1.5 min recalling only words presented in the immediately-preceding list. After 90 s, a tone sounded on the computer; the tone was their signal to draw a line and begin guessing, writing down words that could plausibly have been present in the list. They were told to do their best to generate as many guesses as necessary to fill in all 15 slots on the answer sheet. Note that both free and forced groups performed identical tasks for the first 90 s of the recall period for each list: They were told to recall as many words as possible without guessing. The groups differed only in the instructional set provided for the final 30 s of each recall period.

After receiving the appropriate instructions, subjects were asked whether they had any questions. After any misunderstandings were clarified, they were told that sometimes after they recalled a list and turned over their recall sheet, they would be asked to recall the list again for another 2 min. The experimenter stressed that on each test, subjects should try for the entire recall period, even if they thought they had exhausted their memory for the list. On every test, a tone sounded when subjects had 30 s remaining, at which time they were to draw a line under the last word recalled; the free recall group then
continued to try for 30 s, and the forced recall group began guessing. The experimenter explained to subjects that they would recall each list as many as three times and that a voice on the computer speakers would always tell them when they should be recalling a list.

This first phase of the experiment (the presentation of lists and corresponding initial recall tests) took approximately 1 hour and 10 min. After presentation of the 18 lists and the accompanying recall sessions, subjects received instructions for the final free recall test, accompanied by an explanation of the remember/know distinction.

On the final free recall test, subjects were given 15 min in which to recall all the words they could retrieve from the study lists. As in the initial recall instructions, they were told that they should be confident that everything they wrote down had in fact occurred in the study lists. In addition, they were asked to classify each item they recalled with respect to whether they remembered its original occurrence in the lists or whether they simply knew it had occurred. Detailed instructions and examples were given to ensure that subjects understood the distinction between remembering and knowing. In summary, subjects were told that remember judgments should be reserved for items for which they held a vivid memory of the occurrence of the item on the speakers; they must recollect some specific aspect of the presentation episode to assign this judgment. For example, they were told that if they recollected something distinctive in the speaker's voice when she said the word, or if they recalled the word that came before it or after it, they should place an R (for remember) by the item. If, however, they were certain that an item had occurred on the list but lacked the feeling of being able to mentally re-experience the
presentation episode, they should place a K (for know) beside the item. It was emphasized that the remember/know distinction should be made with respect to the original presentation episode (i.e., subjects should not base the judgment on whether they remembered or knew they had written the item on a recall test). Subjects were encouraged to ask questions to clarify the remember/know distinction.

Finally, it was explained to subjects that at one-minute intervals they would hear a tone signaling them to draw a line under the last item they had produced to allow the experimenter to track their recall progress across the 15 min time period. They were instructed to attempt to recall items for the entire 15 min period.

The entire experimental session took approximately 1 hour and 45 minutes.

Results and Discussion

Initial Tests

Results from the initial tests are displayed in Table 3 as the probability of recalling studied items (at the top) and critical items (at the bottom) as a function of test number, instructional set, and time allotted. These results correspond to the conditions in which subjects took 3 initial tests on lists. The initial recall results when only one test was taken do not differ statistically from the results of test 1 in the 3-test condition and so will not be considered further. (There should have been no difference because the only difference between the two conditions is whether or not other tests followed the initial recall test for a list, and subjects did not know whether other tests would occur until after the first initial test had been completed.)
Table 3. Probability of recall of studied items and critical nonpresented items as a function of test number, instructional set, and time allotted for the initial tests in Experiment 2.

<table>
<thead>
<tr>
<th>Studied Items</th>
<th>Free 90 s</th>
<th>Free 2 min</th>
<th>Forced 90 s</th>
<th>Forced 2 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>.57</td>
<td>.58</td>
<td>.59</td>
<td>.64</td>
</tr>
<tr>
<td>Test 2</td>
<td>.56</td>
<td>.56</td>
<td>.57</td>
<td>.62</td>
</tr>
<tr>
<td>Test 3</td>
<td>.55</td>
<td>.56</td>
<td>.56</td>
<td>.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Critical Items</th>
<th>Free 90 s</th>
<th>Free 2 min</th>
<th>Forced 90 s</th>
<th>Forced 2 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>.43</td>
<td>.43</td>
<td>.49</td>
<td>.79</td>
</tr>
<tr>
<td>Test 2</td>
<td>.44</td>
<td>.44</td>
<td>.53</td>
<td>.77</td>
</tr>
<tr>
<td>Test 3</td>
<td>.45</td>
<td>.45</td>
<td>.54</td>
<td>.79</td>
</tr>
</tbody>
</table>

**Studied Items.** An inspection of the results from the studied items shows that subjects in the free and forced conditions produced a similar number of studied items in the first 90 s of the recall period. This result was as expected, due to the instructions to the forced group to produce guesses in only the last 30 s of the recall period. By the end of the first 90 s of each recall period, both groups had recalled about the same number of items. Furthermore, subjects in the free recall group had recalled about as many items as they would recollect in the entire test period. That is, the extra 30 s of trying to recall produced no benefit for the free recall group; in contrast, the 30 s of guessing by the forced recall group did lead to greater production of studied items. This claim is supported by statistical analyses, in which a 2 (test instructions) x 2 (time allotted) x 3 (test number) ANOVA was performed. There was a reliable effect of time allotted,
\( F(1,62) = 87.50, \text{MSE} = .00 \) and a reliable time x test instruction interaction, \( F(1,62) = 57.87, \text{MSE} = .00 \). This interaction supports the claim that the 30 s spent guessing in the forced recall group led to greater production of studied items, but spending this time attempting to recall items (as done in the free recall condition) did not produce this increase. The main effect of test instruction was not reliable, \( F(1,62) = 2.33, \text{MSE} = .06 \).

There was a reliable effect of test number, \( F(1,124) = 11.45, \text{MSE} = .00 \), but this difference does not reflect hypermnesia; instead, there was forgetting of studied items across tests. As mentioned in the lead-in to this experiment, this result was not entirely unexpected, given the nature of the lists (i.e., short lists that included abstract words).

Returning to the observation that subjects in the forced recall condition produced more studied items in the final 30 s (whereas subjects in the free recall group did not), there is reason to believe that the enhanced production of studied items in the forced recall condition is not accompanied by conscious recollection of their prior presentation, but that production of these items occurs because when instructed to guess, subjects often produced associates to the words that had been retrieved; therefore, subjects produced items that had been presented. If subjects were producing items as guesses on a first test that they came to recognize as having occurred in the list (during the process of producing the items), we would expect that on subsequent tests these items would be produced in the first 90 s (as being items recalled from the list). This did not happen; in fact, the same pattern (of a drop in recall across tests) as seen in the overall analysis occurred in the first 90 s of the forced recall group. That is, recall dropped from .59 on test 1 to .56 on test 3, \( t(31) = 2.83, \text{SEM} = .008 \). Therefore, the increase in studied items produced as a result of
guessing is likely due to the fact that the associative nature of the list made items
“guessable” and not to the possibility that guessing helped subjects remember more items.

In summary, results with respect to the studied items on the initial tests support the
idea that subjects in the forced recall condition were following instructions: They recalled
items from the list in the first 90 s of the recall period, and they produced plausible guesses
as to words that could have been in the list for the last 30 s. There was no evidence that
this process of guessing (even when producing correct responses) enhanced later recall of
the studied items.

Critical Items. Results for the critical items were of more interest in this initial
phase of the experiment. These results appear at the bottom of Table 3. A 2 \times 2 \times 3
ANOVA showed reliable main effects of test instruction, \( F(1,62) = 16.21, \text{MSE} = .26 \), and
test time, \( F(1,62) = 50.10, \text{MSE} = .03 \). More critical items were produced in the forced
recall condition than in the free recall condition, and more critical items occurred after 2
min than after 90 s. In addition, there was a reliable interaction between these 2 factors,
\( F(1,62) = 47.50, \text{MSE} = .03 \); indicating that there was a larger effect of test instruction
after 2 min than after 90 s. This finding was predicted because subjects in the forced
condition were instructed to produce guesses in the last 30 s; in the first 90 s, both groups
were instructed to recall only items they were sure had been presented. Therefore, this
interaction demonstrates that when instructed to produce guesses, subjects often produced
the critical items.

Although the main effect of test number did not reach significance, \( F(2,124) = 1.46, \text{MSE} = .01 \), the interaction between test number and time was reliable, \( F(2,124) = \)
3.84, $\text{MSE} = .00$, as was the 3-way interaction, $F(2,124) = 3.27$, $\text{MSE} = .00$. The latter interaction can be interpreted by saying that although there was no main effect for number of tests taken, this factor did affect performance in the first 90 s of the forced recall condition. This finding makes sense and is line with predictions. If we assume that guessing can harm memory by promoting later confusion between what is remembered and what was previously produced as a guess, we would expect that producing guesses (in the last 30 s of a forced recall test) might later induce people to recall these guesses as having been presented. Inducing subjects to produce the critical item as a guess on test 1 leads them to later “recall” the item in the free recall portion (first 90 s) of the subsequent tests. This interpretation is backed by a reliable difference between the probabilities of erroneous recall on test 1 (.49) and test 3 (.54) in this condition, $t(31) = 2.33$, $\text{SEM} = .02$.

Further evidence for the above interpretation of the interaction can be found by an examination of the individual guesses produced by subjects in the guessing period of the first forced recall test. Of the 57 total critical items produced as guesses on Test 1, 7 of these were then “recalled” on Test 2 (i.e., produced in the first 90 s). These items account completely for the .04 difference between the probability of recall on Test 1 (.49) and Test 2 (.53). Therefore, the conclusion that forced recall can harm later memory by causing source confusions seems firmly grounded.

In summary, the pattern of results manifested for critical items stands in contrast to the pattern of findings with studied items. For studied items, there was a decrease in recall across tests; this decrease did not interact with test instruction or with time allotted to take the test. Subjects forgot items across successive tests. However, for critical items,
although there was no main effect for test number, this variable did exert an effect in that
subjects in the forced recall condition showed an increase in false recall across successive
tests. That is, when subjects were induced to guess on an initial test, testing enhanced
later false recall.

**Final Free Recall**

Results from the final free recall tests are examined below; the overall probabilities
of recall are considered first, followed by an examination of the pattern of subjects’
remember responses.

**Overall.** The mean probabilities of recalling studied items and critical items as a
function of number of initial tests are presented in Table 4. A 2 (item type) x 3 (number of
prior tests) x 2 (instructional set on initial tests) ANOVA showed no effect of item type:
Critical nonpresented items were recalled with a probability indistinguishable from the
probability of recall of studied items, F(1,62) <1. The ANOVA revealed one reliable main
effect: number of prior tests, F(2,124) = 33.55, $\text{MSE} = .03$. Planned comparisons
revealed that for studied items, taking 3 tests enhanced later performance relative to a
single previous test, $t(63) = 8.33$, $\text{SEM} = .01$, and taking a test enhanced performance
relative to the no test condition, $t(63) = 8.14$, $\text{SEM} = .01$. For critical items, there was no
enhancement of false recall for the 3 test condition relative to the 1 test condition, $t(63)<1$;
the 1 test condition, however, did lead to higher levels of false recall than the no test
condition, $t(63) = 2.26$, $\text{SEM} = .03$. Nevertheless, the primary prediction of the
experiment was confirmed: The number of initial tests taken played a critical role in
determining what was later recalled.
Table 4. Probability of recall on the final free recall test of Experiment 2 as a function of item type, number of prior tests, and retrieval orientation on the prior tests.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Retrieval Orientation on Initial Tests</th>
<th>Free</th>
<th>Forced</th>
<th>Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studied Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Prior Tests</td>
<td></td>
<td>.43</td>
<td>.45</td>
<td>.44</td>
</tr>
<tr>
<td>1 Prior Test</td>
<td></td>
<td>.32</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>No Prior Tests</td>
<td></td>
<td>.18</td>
<td>.24</td>
<td>.21</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>.31</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>Critical Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Prior Tests</td>
<td></td>
<td>.39</td>
<td>.35</td>
<td>.37</td>
</tr>
<tr>
<td>1 Prior Test</td>
<td></td>
<td>.33</td>
<td>.37</td>
<td>.35</td>
</tr>
<tr>
<td>No Prior Tests</td>
<td></td>
<td>.25</td>
<td>.30</td>
<td>.28</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>.32</td>
<td>.34</td>
<td></td>
</tr>
</tbody>
</table>

The ANOVA also revealed a reliable interaction between number of prior tests and item type, $F(2, 124) = 8.86$, $MSE = .02$. This interaction can be interpreted as showing that testing had a larger effect on studied items than on critical items. Collapsing across the free and forced recall conditions, there was a .23 difference in the probability of recalling an item in the no test and 3 test conditions for studied items; the corresponding difference for critical items was .095.
It is interesting to note that this difference between studied and critical items is due in part to the fact that critical items are recalled with a high probability even in the absence of prior tests: In the no test condition, the probability of recalling a critical item exceeded the probability of recalling a studied items, $t(64) = 2.35$, $\text{SEM} = .026$ (collapsing across the test instruction variable). Indeed, another way of interpreting the test number $x$ item type interaction is to say that the effect of item type differed across number of initial tests taken. As shown above, the probability of recall of critical items exceeded that of studied items in the no test condition; the opposite pattern was true after 3 tests, $t(63) = 2.30$, $\text{SEM} = .03$.

The main effect of instructions given on the initial test was not reliable, $F(1,62) < 1$, nor did this factor enter into any interactions. Therefore, the differences observed in the forced and free conditions on the initial tests did not carry over onto the final free recall test. There was, however, only a limited effect of this variable even on the initial tests, so perhaps it is not too surprising that this variable did not affect final free recall. I had predicted, however, that the critical intrusion rate on the final free recall test would be for the group that was induced to guess on initial tests than for the group that received standard, free recall instructions.

Remember judgments. Table 5 displays the proportion of items for which subjects claimed to remember the specific instance of presentation. Therefore, these data represent a subset of the data occurring in Table 4. An inspection of Table 5 shows that the overall pattern of results looks very similar to the overall recall results. However, subjects did not simply classify most of their recalled items as old. Of the studied items recalled, .63
(.205/.325) of them were classified as *remembered*; the corresponding proportion for critical items was .48.

Table 5. Probability of obtaining a *remember* judgment on the final free recall test of Experiment 2 as a function of item type, number of prior tests, and retrieval orientation on the prior tests.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Retrieval Orientation on Initial Tests</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Free</td>
<td>Forced</td>
<td>Means</td>
</tr>
<tr>
<td>Studied Items</td>
<td></td>
<td>.28</td>
<td>.27</td>
<td>.28</td>
</tr>
<tr>
<td>3 Prior Tests</td>
<td></td>
<td>.20</td>
<td>.20</td>
<td>.20</td>
</tr>
<tr>
<td>1 Prior Test</td>
<td></td>
<td>.12</td>
<td>.16</td>
<td>.14</td>
</tr>
<tr>
<td>No Prior Tests</td>
<td></td>
<td>.20</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>.20</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Critical Items</td>
<td></td>
<td>.23</td>
<td>.20</td>
<td>.22</td>
</tr>
<tr>
<td>3 Prior Tests</td>
<td></td>
<td>.11</td>
<td>.15</td>
<td>.13</td>
</tr>
<tr>
<td>1 Prior Test</td>
<td></td>
<td>.10</td>
<td>.16</td>
<td>.13</td>
</tr>
<tr>
<td>No Prior Tests</td>
<td></td>
<td>.15</td>
<td>.17</td>
<td></td>
</tr>
</tbody>
</table>

A 2 x 2 x 3 ANOVA detected a reliable effect of item type: Studied items were *remembered* with a higher probability than critical items, $F(1,62) = 6.80$, $MSE = .03$. Of course, this finding makes sense in that there was no presentation to be *remembered* for the critical items. (Roediger and McDermott, 1995, however, did achieve equivalent levels of *remember* judgments on a recognition test.) Nevertheless, although critical items
received fewer *remember* responses than studied items, critical items were often labeled as being *remembered* ($M = .16$, collapsing across the other conditions). Another way of considering this result is that 48% of the critical items that were recalled were classified as being *remembered* (i.e., $16/.33$). A comparable analysis of studied items shows that 63% of the recalled items were classified as being *remembered*. This way of examining the results is interesting, in part, because it shows that subjects do not simply classify every recalled item as being *remembered*. Therefore, it suggests that the assumption that recall primarily measures the phenomenological experience of *remembering* is ill-founded; familiarity seems to play a substantial role in recall, at least when it occurs after a delay, as in this experiment. (Presumably, if subjects were asked to classify the items recalled on an immediate free recall test as being *remembered* or *known*, almost all would be *remembered*, at least with the short lists used here.)

The primary question in examining *remember* responses was to determine whether taking a test might enhance the probability with which people would later claim to *remember* an item. There was a reliable effect of number of prior tests, $F(2,124) = 25.84$, $MSE = .02$. Furthermore, planned comparisons demonstrated that relative to a single prior test, taking 3 prior tests enhanced the later phenomenological experience of *remembering* both for studied items, $t(31) = 3.72$, $SEM = .022$ for the free recall condition and $t(31) = 4.52$, $SEM = .015$ for the forced recall condition, and for critical items, $t(31) = 2.73$, $SEM = .015$ and $t(31) = 1.82$, $SEM = .03$ for the free and forced conditions, respectively. Note that this pattern occurred despite the explicit instructions to subjects that they were to make the *remember*/*know* judgments with respect to whether
they could recollect the initial presentation of the item, and not with respect to whether they recollected having produced the item on a recall test.

As in the overall results, the interaction between item type and number of prior tests was reliable for the *remember* judgments, $F(2, 124) = 4.15$, $MSE = .01$. Taking an initial test enhanced the later phenomenological experience of *remembering* of studied items more than it affected critical items.

No other main effects or interactions were reliable. As in the overall results, there was no main effect of the forced/free instructional manipulation on the initial tests. In summary, the basic pattern of *remember* judgments mirrored the overall recall data. Subjects produced critical items at a high rate, they often claimed to be able to recollect the instance of presentation of these items, and this erroneous recollection was enhanced by taking prior tests.

The results of the first two experiments demonstrate that false memories can be quite robust. They occur with a high probability and are often accompanied by a feeling of being able to mentally re-live the presentation episode of the non-occurring event. Experiment 3 was designed to explore more directly the persistence of this false recall effect; specifically, the question was whether such false recall could withstand repeated presentation of the study lists.

**Experiment 3**

Experiment 3 was designed to determine whether false recall in the paradigm under investigation could be attenuated or eliminated across multiple study/test trials. Because correct recall would have approached ceiling quickly using only 15 words per list, three
15-word associative sets were presented in each study list. That is, subjects studied one long list of 45 words, which consisted of three groups of 15 associates (e.g., hot, snow, warm...bed, rest, awake... thread, pin, eye....). Items were presented blocked by associative set for half the subjects and in a random order for the other half.

Assuming these lists function like categorized lists, previous research would indicate that for the studied items, there should be a main effect for ordering, with blocked lists producing higher levels of recall than randomly-ordered lists (Dallett, 1964). One question addressed in this experiment was whether there would also be an effect of the blocked/random variable for critical nonpresented items. No firm predictions could be made for this variable because a plausible rationale exists for predicting an effect in either direction. It could be that in trying to remember the list items (either during encoding or retrieval or both), subjects would try to categorize and organize the lists according to common themes within the lists. Obviously, the "common themes" may be represented by the critical nonpresented items. Because items from the same "theme" are scattered throughout the list in the case of random presentation, subjects might be more confused about which specific items occurred on the random lists than on the blocked lists. If so, they might be expected to intrude the critical nonpresented words more often in random than in blocked lists. Conversely, one could argue that blocked presentation should produce higher false recall levels than random presentation because blocking words might encourage more relational processing of the words and therefore be more likely to elicit the critical word as an implicit associative response (Underwood, 1965) or to induce subjects to assume that the word had been presented. In this analysis, the theme is easier
to perceive in the blocked case. Therefore, as stated above, this variable was an exploreatory one. An additional question was whether any blocked/random effect (or lack thereof) for critical items would change across trials.

The primary question of the experiment was whether the proportion of critical items recalled would decrease or disappear across trials. A plausible hypothesis was that a decrease would occur because as subjects continue to learn the items present in the list, they may also learn the items not present (or conspicuous in their absence). For example, subjects may produce a critical nonpresented word on a first test but be unsure whether it had actually occurred. They could check for this word during the second presentation. Another interesting possibility was that if subjects recalled a critical item once, they might continue to recall it, carrying over their error despite repeated study opportunities. Kay (1955) reported this pattern in recall of prose passages. He presented subjects with the passage and then tested them after a short (5 min) delay. Following the test, he presented the passage again, waited 1 week, tested subjects again, and re-presented the passage. This sequence was carried out until six recall tests had occurred. Kay found that subjects had great difficulty editing out intrusions, despite the fact that the original, correct material had been encountered more recently than the erroneous recall. (Kay’s, 1955, general finding was later replicated by Howe, 1970, and extended to false recognition, Howe, 1972.)

Kay (1955) used long (i.e., 1 week) delays between successive re-presentations of the passage and the test phases because he assumed long delays were necessary to minimize the influence of the intervening study episodes, thereby allowing intrusions to
persist across test trials despite the opportunity to re-encounter the study material. Because the false recall effect on immediate tests is very robust with the paradigm used here, I decided to give the tests immediately following list re-presentation to determine whether the intrusions might persist even under conditions that seem likely to promote their disappearance.

Method

Subjects

Forty Rice University students participated in this experiment in exchange for course credit or pay.

Design

A 2 (Ordering: Blocked, Random) x 6 (Test Session: Trials 1-5, Day 2) x 2 (Item Type: Studied, Critical) mixed design was used, with ordering of lists manipulated between subjects and the other two variables instantiated within subjects.

Materials

Six of the 15-word lists used in Roediger and McDermott’s (1995, Experiment 2) report were selected on the basis of their producing relatively high levels of recall of the critical items. These lists were divided into two 45-word lists of three 15-word associative sets per list. One list contained associates of the 3 critical items cold, sleep, and needle, and the other list contained associates of fruit, chair, and mountain. Each subject received one 45-word list. Items within lists occurred in a fixed order, and the same order was used on all five study presentations. For lists presented in a blocked fashion, sets of associates were arbitrarily assigned to positions within the list. For lists presented
randomly, items were randomly assigned to 45 positions within the list with the stipulation that no more than three items from one associative set could occur in succession.

For example, in the blocked condition, one list contained the 15 associates to cold, followed by the 15 associates to sleep, followed by the 15 associates to needle. For random presentation, items from the 3 associative sets occurred randomly (in a fixed order) within the list. There were no demarcations between associative sets within lists, and subjects were not informed about the relations of words within the lists in either condition.

Lists were recorded digitally in a female voice in stereo using a 22050 Hz sampling rate and 16-bit resolution. Words occurred at the approximate rate of 1 word per sec, and they were presented via LabTec CS 1400 stereo speakers. Subjects were provided with 8 1/2" x 11" sheets of paper on which to record their responses.

Procedure

Subjects were tested in groups of 4 or fewer. They were told that they were participating in an experiment designed to examine their ability to learn lists of words. It was explained that they would hear a list of 45 words presented on computer speakers, the list would be presented 5 times, and each presentation would be followed by a recall test. They were told that they would hear the same list every time and that the ordering of words within the list would remain constant. They were told to recall the list by writing down as many words as they could remember with the stipulation that they were to be confident that every word they wrote had been present in the list. They were given 4 min in which to recall the list, and the end of the recall period was indicated by a tone on the
computer. At this time, subjects turned over their test sheet and waited for the next presentation. Questions were answered, and list presentation was begun.

Upon returning a day later, subjects were instructed to recall as many items as possible from the list they had learned in the previous session. As on the initial tests, they were explicitly warned against guessing. Subjects were given 12 min in which to recall, and a tone sounded on the computer every minute, signaling subjects to draw a line under the last item recalled and continue to try to remember more items. Questions were answered, and the recall period began. Note that there was no presentation of the list on the second day.

Results and Discussion

Recall probabilities (both accurate and erroneous) are portrayed in Figure 7. Results from the first day of the experiment (the multitrial learning session) are discussed first, followed by an examination of the final recall test given a day later. Finally, conditional recall probabilities are considered.

Day 1

An overall ANOVA indicated that there was no reliable 3-way interaction among item type (studied, critical), study order (blocked, random), and test number, $F(4,152) < 1$. Because the focus of the experiment was on critical items, the data for studied and critical items were not collapsed; instead, separate 2-way analyses were performed for studied items and critical items to examine the effects of study order and test number.

Studied Items. Initial observations of the top two curves in the figure suggest that the predicted pattern of effects occurred for studied items: Recall increased across trials,
demonstrating learning of the list, and performance was generally better in the blocked condition than in the random condition. However, ceiling effects may have masked the blocked/random effect on the last couple of trials.

Figure 7. Probability of recall of studied items (solid lines) and critical nonpresented items (dashed lines) as a function of test trial and presentation order for Experiment 3.

A 2 (ordering) x 5 (test session) mixed ANOVA on the proportions correct was consistent with these observations. A main effect for test session was found, $F(4,152) = 826.87$, $\text{MSE} = .00$, demonstrating learning across trials. The blocked/random variable had no reliable effect, $F(1,38) = 1.87$, $\text{MSE} = .05$, but the interaction between test session and list order was reliable, $F(4,152) = 7.17$, $\text{MSE} = .002$, consistent with the claim that a ceiling effect obscured the ordering effect in the last few trials.

Critical Items. There were two main questions with respect to critical items in this experiment: (1) To what extent would production levels of critical items decrease across trials? and (2) Would there be a blocked/random effect? An examination of Figure 7
suggests that the probability of false recall was greater in the blocked condition than in the random condition. False recall also diminished across trials, with proportion of intrusions dropping between Trials 1 and 5 from .57 to .32 for the blocked condition and from .30 to .20 for the random condition. Nevertheless, recall of the critical items was clearly not eliminated, as subjects produced substantial levels of false recall, even after hearing 5 presentations of the list.

A 2 (ordering) x 5 (trial) ANOVA on the proportions generally confirmed these observations. There was a significant main effect of trial number, $F(4,152) = 4.31$, MSE = .05. However, the effect of study order did not reach (but did approach) the criterion for significance, $F(1,38) = 3.54$, MSE = .38, $p = .068$. There was no significant interaction $F(4,152) < 1$. Thus, false recall in both the blocked and the random condition decreased at a similar rate across trials. Whether list ordering exerts an effect on false recall is inconclusive if one considers this experiment alone. However, Toglia and his colleagues have recently reported a reliable effect of this variable in the same paradigm, both on an immediate test (Toglia, Neuschatz, Goodwin, & Lyon, 1995a) and after a 1-week delay (Toglia, Neuschatz, Goodwin, & Lyon, 1995b). Therefore, given the findings of Toglia et al. (1995a, 1995b), combined with the marginally reliable effect obtained here, it seems safe to conclude that greater false recall occurs after blocked than after random conditions.

Finally, it is interesting to note that on the first trial in the blocked condition, the proportion of critical nonpresented items recalled (.57) greatly exceeded the proportion of studied items recalled (.38), $t(19) = 2.39$, SEM = .079. By Trial 3, however, this pattern
had reversed, showing a reliable advantage for studied items, \( t(19) = 2.8, \text{ SEM } = .089 \).

**Day 2**

The data points at the far right side of Figure 7 represent performance on the final free recall test, which occurred one day later, in the absence of an intervening study opportunity. These data suggest that recall of studied items decreased after one day (i.e., items were forgotten), whereas intrusion of critical nonpresented items increased after a day. Paired t-tests confirmed these observations. For studied items, the drop in recall levels between Trial 5 of Day 1 (.84) and Day 2 (.80) was reliable, \( t(39) = 2.81, \text{ SEM } = .012 \). Conversely, the level of false recall rose between Trials 5 (.26) and Day 2 (.33), \( t(39) = 2.08, \text{ SEM } = .032 \).

An examination of noncritical intrusions showed that subjects rarely made such errors. Collapsing across trial number and study order, the mean number of noncritical intrusions was .49 items per subject. Thus, subjects were not simply guessing when instructed to recall the lists.

**Conditionalized recall probabilities**

The data for the critical items in this experiment are interesting in part because subjects had difficulty correcting their errors despite repeated opportunities to do so. However, given the data reported thus far, an alternative interpretation exists; that is, perhaps subjects successfully edited out their initial intrusions on the first few trials, but new errors arose later in the session. Recall that each subject received a list centered around 3 critical items. Suppose subjects generally recalled 2 of these items on Trial 1 and (say) by Trial 3 had learned not to include these 2 but then began to (erroneously) recall
the third critical item. This scenario (or a similar one) would produce the obtained pattern of results.

To determine whether such a pattern might underlie these data, I examined new intrusions on each test trial. Shown in Figure 8 are the overall false recall levels for the blocked and random conditions (repeated from Figure 7), along with the accompanying proportions of new intrusions recalled on each trial. "New" intrusions were defined as those that had not appeared on any previous test trial. It is evident from this graph that with the exception of the second trial, subjects rarely introduced new intrusions into their protocols; the learning curves are not being artificially inflated by new intrusions occurring in later trials. Therefore, the conclusion that subjects have difficulty correcting their errors seems firmly grounded.

Figure 8. Overall probability of recall of critical nonpresented items (dashed lines) and the probability of recalling "new" critical items as a function of test trial and presentation order for Experiment 3.
In summary, results of Experiment 3 demonstrate that when given repeated opportunities to hear the study list, subjects could reduce the proportion of critical nonpresented items recalled. However, they did not eliminate these errors. An analysis of conditionalized recall probabilities showed that this latter finding was primarily attributable to subjects' carrying over their initial errors onto later tests, and not to the introduction of new intrusions.

General Discussion

A number of noteworthy results were obtained in the three experiments reported here.

- First, I obtained high levels of false recall following presentation of lists of associated words, thereby replicating Deese's (1959b) and Roediger and McDermott's (1995) findings (see also Payne et al., 1996; Read, 1996; and Schacter et al., 1996). The intrusions were obtained despite an explicit warning to subjects against guessing.

- More interestingly, the high levels of false recall were maintained over a short delay (30 s) and remained strong two days later. Indeed, in Experiment 1, after two days, the probability of false recall of critical nonpresented items exceeded the probability of recall of items actually presented.

- Both accurate recall and false recall were increased on a final free recall test by a single prior testing of the list; this effect persisted across a 2 day delay in Experiment 1 and was extrapolated in Experiment 2 to show that three tests produce a greater effect on both accurate remembering and false remembering than does a single test.

- Encouraging subjects to guess on an initial test in the forced recall condition of
Experiment 2 caused them to "recall" their guesses (of critical items) on immediately succeeding tests as having been presented. Therefore, a phenomenon analogous to hypermnnesia was obtained with a forced recall procedure, even though no hypermnnesia was obtained for studied items.

- When Tulving's (1985) remember/know procedure was employed on a final free recall test (in Experiment 2), subjects often claimed to remember the presentation of nonpresented items. Furthermore, this error was exacerbated by multiple prior tests: When 3 prior tests had been taken, the probability that subjects would claim to remember the presentation of the critical item exceeded the probability of remembering the item when only 1 prior test had occurred. Likewise, taking a prior test enhanced the later (erroneous) subjective feeling of remembering relative to the no test condition. Thus, testing was found to exert a powerful influence on later mis-remembering.

- Although robust levels of remember responses were obtained in Experiment 2 for both veridical and false recall, by no means did subjects classify all of their responses as being remembered. This finding seems to suggest that perhaps free recall (at least after some delay) is not a pure measure of recollective experience.

- In a multtrial learning paradigm, recall of the critical items dropped over trials but stabilized at 20% or 32% (depending on blocked or random study condition), indicating that subjects could not completely edit the erroneous responses even after five study/test cycles. An examination of conditionalized recall probabilities of the intrusions supported this claim. This finding provides converging evidence for the persistence of illusory memories in that they are resistant to correction.
• Blocked presentation enhanced accurate recall relative to random presentation (for early trials); the finding for critical items was less clear-cut, as the blocked/random difference fell short of the criterion for significance. However, given that Toglia et al. (1995a, 1995b) obtained a robust blocked/random difference in a similar paradigm, the effect for critical items likely parallels the effect for studied items.

The present results speak to several issues in the development and maintenance of false memories. I discuss in turn how these results bear on three issues: (1) encoding factors that encourage development of false memories; (2) the critical role of retrieval factors in enhancing false memories; and (3) the persistence of the illusory memories created in this paradigm.

**Encoding factors**

Roediger and McDermott (1995) noted that the earliest idea used to explain false recognition in Underwood's (1965) paradigm also provides a workable account of the more recent findings. Underwood (1965) proposed that an implicit associative response occurring during encoding (see or hear *thread*, and then think *needle*) could explain why subjects would later falsely recognize *needle* as having occurred in the list. In the paradigm used here, subjects heard 15 words derived from a common nonpresented word. The nonpresented word is in turn an associate of many of the list words and therefore may be activated during list presentation. If the activation results in conscious awareness of the nonpresented word, then this might explain why, when Tulving's (1985) *remember/know* procedure is used, subjects claim to *remember* the occurrence of the critical nonpresented words (Roediger & McDermott, 1995, Experiment 2; Payne et al.,
1996; Schacter et al., 1996). The difficulty would resemble that of reality monitoring, or internal source monitoring (Johnson & Raye, 1981; Johnson, et al., 1993), in which subjects must judge whether events they recognize were encountered externally or were internally generated. As Johnson and her colleagues have demonstrated, subjects often confuse imagined events with experienced events (Johnson, 1995).

The output position curves of the initial tests in Experiment 1 are fully consistent with this explanation. Prior results reported by Roediger and McDermott (1995) (and replicated in Experiment 1 here as well as by Payne et al., 1996) showed that when a free recall test follows immediately after list presentation, subjects tend to recall the critical nonpresented item toward the end of their output (see Figure 4). These results might seem to suggest that in the process of recalling many studied items, the critical items occur to subjects and are therefore written down. That is, perhaps the phenomenon is retrieval-induced. However, the results of the 30 s delay condition of Experiment 1 seem to suggest that this is not the case. When the primary memory component of recall is eliminated (by introduction of a 30 s filled delay), there is no systematic placement of the item in the recall protocol. These results are consistent with the possibility that during the course of hearing the study list, subjects think of the critical item.

Regardless of whether the phenomenon originates at encoding, it can clearly be affected by encoding factors. Blocked presentation enhances recall of both correct items and critical nonpresented items (see Experiment 3 reported here; Toglia, et al., 1995a, 1995b). In addition, Toglia et al. (1995a) varied level of processing during encoding and showed that false recall (like correct recall) occurs more frequently after deep, semantic
processing (a pleasantness rating task) than after shallow, graphemic processing (determining whether each word contained an “a”). Both manipulations of list structure (blocked/random) and level of processing (graphemic, semantic) show that encoding factors affect later probability of false recall. Perhaps it is not too surprising that semantic analysis of meaningfully related words enhances false recall relative to graphemic analysis. More surprising is the outcome that blocked presentation leads to greater false recall than does random presentation. As mentioned in the lead-in to Experiment 3, one might hypothesize that repeated blocked presentation as in Experiment 3 would cause subjects to identify the nature of the list and more easily discern that a particular item they were expecting to hear had not occurred in the list. However, this did not happen. It would be interesting to see if the same pattern would occur if subjects received simultaneous visual presentation of the list, which would make the list structure more readily apparent (e.g., Elmes, Roediger, Wilkinson & Greener, 1972).

The results of Experiment 3 can also be interpreted within this encoding framework: If the associated items are presented together at study, then the context may make each presented item more likely to arouse the common associate than if the relevant items are dispersed throughout the list. Similarly, semantic analysis of list items would be more likely to lead to activation of links among associates than would shallower, graphemic analysis. Although the investigation into how study manipulations affect this memory illusion is just beginning, the present results and those of Toglia et al. (1995a, 1995b) fit comfortably into the ideas first proposed by Underwood (1965). However, other results indicate that encoding factors are only partly responsible for these effects.
Retrieval factors, considered next, also play a large role.

**Retrieval factors**

The study of the effect of taking an initial test on performance on a later test has a long history (e.g., Ballard, 1913; Spitzer, 1939). One general finding is that repeated testing can lead to greater overall recall of the target material under certain conditions, a finding called hypermnnesia (Erdelyi & Becker, 1974; see Payne, 1987 for a review). A second consistent finding is the testing effect: Taking a test leads to greater recall or recognition on a later test than found in a control condition in which no first test is given (e.g., Spitzer, 1939; Wheeler & Roediger, 1992). However, the effect of testing can also have deleterious consequences on later recall, as seen in these experiments (see too Schooler, Foster, & Loftus, 1988; Payne et al., 1996, Experiment 2). That is, the testing effect exists for false recall, as well as for accurate recall; the present experiments show that taking a first test enhances later false recall just as it does veridical recall (see Figure 5). The act of recall presumably provides retrieval practice for recalled items and makes them more accessible on a later test. Indeed, many experiments have shown that recall of events produces greater facilitation on a later test than does the actual re-presentation of the events (Hogan & Kintsch, 1971).

The results of Experiments 1 and 2 demonstrate that the same facilitation occurs for false recall as for accurate recall (see Roediger, Jacoby, & McDermott, 1996 for further evidence on this point). The impact of initial testing on a later recognition test, however, is less clearcut. Roediger and McDermott (1995, Experiment 2) found testing effects for studied items and critical nonpresented items on a recognition test. However,
using the same general paradigm, Schacter et al. (1996) failed to find reliable testing
effects for studied items or critical items. In addition, in an experiment that was essentially
failed to find a testing effect for critical items, although the effect was present for studied
items. The reason for the discrepant findings within the recognition tests is unclear.

An examination of 11 experiments that vary somewhat with respect to their
similarity to the original experiment reported by Roediger and McDermott (1995) seems
generally to confirm the existence of a testing effect for veridical recognition; the story for
false recognition, however, is less clear. Table 6 displays the hit rates and critical false
alarm rates as a function of prior testing history for 11 experiments, all of which
conformed fairly closely to the procedures used by Roediger and McDermott (1995).
Several independent variables, in addition to initial test condition, were manipulated in
these experiments. Payne, et al. (1996) varied the retention interval between the initial
study-test phases and the final recognition test; the interval was either 2 min or 24 hours.
Two reports examined individual differences in the false recognition effect: Norman and
Schacter (1996) compared young and older adults, whereas Schacter, et al. (1996)
compared amnesic patients to age-matched control subjects. Although a few data points
in Table 6 fail to show a benefit from prior recall on recognition, the general finding is that
although many variables (e.g., retention interval and age) affect the levels of recognition,
there is a consistent numerical advantage in recognition favoring the conditions in which
there had been an initial test (relative to the conditions in which no initial test had been
administered). As can be derived from Table 6, in 17 of 22 relevant comparisons between
Table 6. Testing effects in recognition memory: Recognition results from studies conforming to Roediger and McDermott's (1995, Experiment 2) procedure as a function of item type and whether an initial free recall test was taken.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Test Conditions</th>
<th>Initial Test?</th>
<th>Item Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Roediger &amp; McDermott (1996, Exp’t 2)</td>
<td>Immediate</td>
<td>Yes</td>
<td>.79</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.65</td>
<td>.72</td>
</tr>
<tr>
<td>Payne, et al. (1996, Exp’t 1)</td>
<td>Immediate</td>
<td>Yes</td>
<td>.73</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.62</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>24 hour delay</td>
<td>Yes</td>
<td>.64</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.47</td>
<td>.62</td>
</tr>
<tr>
<td>Norman &amp; Schacter (1996, Exp’t 1)</td>
<td>Young subjects, no justification*</td>
<td>Yes</td>
<td>.77</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.81</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>Young subjects, justification</td>
<td>Yes</td>
<td>.84</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.75</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>Old subjects, no justification</td>
<td>Yes</td>
<td>.75</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.72</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td>Old subjects, justification</td>
<td>Yes</td>
<td>.82</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.71</td>
<td>.85</td>
</tr>
<tr>
<td>Norman &amp; Schacter (1996, Exp’t 2)</td>
<td>Young subjects</td>
<td>Yes</td>
<td>.79</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.69</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>Old subjects</td>
<td>Yes</td>
<td>.71</td>
<td>.70</td>
</tr>
</tbody>
</table>
Schacter et al. (1996)

<table>
<thead>
<tr>
<th>Amnesic Subjects</th>
<th>Yes</th>
<th>.54</th>
<th>.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Subjects</td>
<td>Yes</td>
<td>.85</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.83</td>
<td>.89</td>
</tr>
<tr>
<td>Mean (collapsing across exp'ts)</td>
<td>Yes</td>
<td>.75</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.67</td>
<td>.70</td>
</tr>
</tbody>
</table>

*Note: Subjects were required in some conditions to justify their *remember* responses by telling the experimenter what specifically they recollected about the presentation episode.

The conditions in which there was an initial test and those in which no initial test occurred, there is a numerical advantage for the condition in which a prior test was taken, with one tie (significant by a one-tailed sign test, p < .05). Collapsing across all experiments, the mean probabilities of recognizing previously tested items were .75 and .73 for studied and critical items, respectively; the comparable recognition rates for lists not previously tested were .67 and .70. When subjected to Wilcoxon signed-ranks tests, these data show a reliable overall effect for studied items (T=3) but not for critical items (T=14.5). Therefore, the evidence seems to converge upon the conclusion that taking an initial test enhances the probability with which people will later classify the studied items as having been presented. Evidence for critical items is less certain.

An analysis of the testing effect with respect to *remember* judgments tells a similar story across experiments. Nine of the experimental conditions reported in Table 6 also employed the *remember/know* procedure. Results from the *remember* conditions of these
Table 7. Testing effects in *Remember* Responses: Results for studies conforming to Roediger and McDermott's (1995, Experiment 2) procedure as a function of item type and whether an initial free recall test was taken.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Test Conditions</th>
<th>Initial Test</th>
<th>Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roediger &amp; McDermott (1996, Exp't 2)</td>
<td>Immediate</td>
<td>Yes</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.41</td>
</tr>
<tr>
<td>Payne, et al. (1996, Exp't 1)</td>
<td>Immediate</td>
<td>Yes</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>24 hour delay</td>
<td>Yes</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.28</td>
</tr>
<tr>
<td>Norman &amp; Schacter (1996, Exp't 1)</td>
<td>Young subjects, no justification*</td>
<td>Yes</td>
<td>.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>Young subjects, justification</td>
<td>Yes</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>Old subjects, no justification</td>
<td>Yes</td>
<td>.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>Old subjects, justification</td>
<td>Yes</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.50</td>
</tr>
<tr>
<td>Schacter et al. (1996)</td>
<td>Amnesic Subjects</td>
<td>Yes</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>Control Subjects</td>
<td>Yes</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>.55</td>
<td>.52</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Means (collapsing across exp'ts)</td>
<td>No</td>
<td>.44</td>
<td>.46</td>
</tr>
</tbody>
</table>

*Note: Subjects were required in some conditions to justify their remember responses by telling the experimenter what specifically they recollected about the presentation episode.

Experiments are presented in Table 7, with 14 out of 18 showing a positive effect of prior testing on remember judgments. For studied items, all of the 9 relevant comparisons showed a numerical advantage for the condition in which an initial test had occurred (relative to the no test condition). The comparable comparison for critical items shows 6 of the 9 conditions revealing a numerical testing effect. Collapsing across all 9 experiments, when a prior recall test had occurred, the probabilities of remember responses were .55 and .52 for studied and critical items, respectively; in the absence of a prior test, the corresponding probabilities were .44 and .46 for studied and critical items, respectively. Wilcoxon signed-ranks tests showed a reliable testing effect for studied items, T=0, although the value for critical items fell just short of significance, T=8.5. In the aggregate, the results with respect to remember judgments bolster the claims made by Roediger and McDermott (1995) that prior testing enhances the later recollection (as manifested in remember judgments) of the studied items; the effect on critical nonstudied items is less conclusive.

In Experiments 1 and 2 reported here, using final free recall, robust testing effects were obtained for both studied and critical items. Perhaps it is not too surprising that a
testing effect for false memories is easier to obtain on a free recall test than a recognition test. This idea is based on the observation that for veridical recall, testing effects are robust on recall tests but more difficult to obtain on recognition. The crucial point in the current experiments, though, is that in the paradigm used here, initial testing affects later false recall in much the same way that it enhances veridical recall.

Contrary to predictions, no hypermnesia-like effect was observed for critical items unless subjects were induced to guess (via forced recall instructions). While Experiment 2 was being conducted, Payne et al. (1996) reported obtaining such an improvement in a variation of the paradigm used in Experiment 2. Payne et al. (1996), however, presented subjects with 6 sets of 10 lists (similar to the blocked condition of Experiment 3 reported here), followed by 3 successive free recall tests. Using this procedure, these researchers were able to observe an increase in production of the critical nonpresented items across tests (.27 on test 1 and .35 on test 3).

There appear to be two major (related) differences in the materials used in their experiments and Experiment 2 reported here: (1) Payne, et al. (1996) presented multiple sets of associates followed by several recall tests, whereas in Experiment 2, subjects took tests after each single set of associates; (2) Payne, et al. (1996) presented subjects with lists that were much longer (60 words) than the 15-word lists used in Experiment 2. Because hypermnesia is usually not obtained (for studied items) without long lists, it may be that long lists are a necessary condition for observing an analogous effect in false recall. It also seems likely that list structure might influence the probability of obtaining hypermnesia; presentation of multiple sets of associates resulted in very high false recall
levels in Experiment 3 reported here. (On trial 1 of the blocked condition of Experiment 3, the probability of accurate recall was .38 and the probability of erroneous recall .57: Combining sets of associates seems to produce very robust levels of false recall.) Resolution of the relative contribution of the 2 factors (list length and list structure), as well as resolution of the underlying question of when successive tests will induce enhanced false recall, must await future research.

An interesting finding emerged in Experiment 3 from a comparison of the last recall trial on Day 1 and the delayed test on Day 2. For both blocked and random conditions, subjects showed a drop in accurate recall but an absolute increase in false recall over time and repeated testing. This finding is reminiscent of the outcome of Posner and Keele’s (1970) study of learning to classify dot patterns. Over delays subjects became less efficient in classifying the actual patterns that had been presented but still retained the ability to quickly classify the prototype from which the patterns had been generated (see Solso & McCarthy, 1981; Spiro, 1980; and Sulin & Dooling, 1974, for similar observations). The present finding, if confirmed by later work, would show absolute increases in false recall over a delay, consistent with Bartlett’s (1932) anecdotal results and speculations.

Research in other paradigms also shows the importance of retrieval in creating false memories. Roediger, et al. (1993) reported an experiment in which subjects studied 60 items and then either took a free recall test or were forced to guess on an initial test (i.e., to produce 60 items that plausibly could have been presented). When asked to judge whether each item they had written actually had been presented on the list, subjects in the
forced recall condition often claimed to recognize their intrusions as having been studied, whereas subjects in the free recall condition did not show this effect. Thus, the process of generating responses can potentially create memories for the presentation of these items. In Experiment 2, critical items produced as guesses (in the forced recall condition) had the tendency to be recalled as veridical memories on subsequent tests. Relatedly, Hyman, Husband, and Billings (1995) reported experiments showing that repeatedly thinking about whether a non-occurring childhood event had happened increased the likelihood that subjects believed that it actually had happened (see too Hyman & Pentland, 1996). These findings converge on the results reported here in demonstrating that the act of retrieval affects not only veridical retention but also false recall. Research in this area is new, but it seems likely from these first experiments that retrieval factors play a critical role in the development and maintenance of false memories.

**Persistence**

The experiments reported here demonstrate that intrusions in free recall tests can be persistent across time and across multiple opportunities for subjects to realize their errors. In Experiment 1 (as well as in the papers by Roediger & McDermott, 1995, Payne et al, 1996, and Schacter et al., 1996), it was shown that on an immediate recall test, probability of recall of the critical nonpresented item approximated the probability of recall of items that had been presented in the middle portion of the list, or items usually thought to represent retrieval from long term memory. After a 2-day delay, false recall probabilities exceeded recall probabilities from all list positions except the first few (see Figure 5). Thus, the critical nonpresented items appear to have been retrieved from long-
term memory much like presented items and, curiously, the "forgetting" of nonpresented events seems to have occurred more slowly than for events that did occur. Indeed, in Experiment 2, in the absence of prior testing, the probability of recalling a critical nonpresented item exceeded the probability of correct recall on a final free recall test. In addition, in Experiment 3, the probability of false recall increased over a one-day delay, whereas the probability of recall of list items decreased.

Experiment 3 was designed to examine directly the persistence of memory illusions obtained in this paradigm by providing subjects with multiple study (and test) opportunities. Although subjects were able to decrease the number of false recalls across trials, they were not able to eliminate them completely. In the blocked condition, the probability of false recall greatly exceeded that of veridical recall on the first trial, but by the second trial, this pattern disappeared. In both blocked and random conditions, recall of list items appeared to be near asymptote by Trials 4 and 5, but the false recall probability did not drop below .32 and .20, respectively. Of course, whether additional study/test trials would cause the drop to continue remains to be seen, but clearly the illusion persists under the conditions used in Experiment 3 here.

Perhaps one reason that false recall was not eliminated in Experiment 3 is that the repeated testing aspect of the procedure somewhat counteracted the effect of repeated study presentations. That is, if recall tests serve to enhance procedures used in retrieval for both false and accurate recall, subjects might not have been able to use effectively the subsequent study presentations to eliminate errors. If subjects received only repeated study presentations (without intervening tests) they may be less likely to produce
intrusions than in the customary study/test procedure.

**Concluding Remarks**

In conclusion, the experiments reported here demonstrate robust memory illusions in which nonpresented items are retrieved from long term memory with a probability greater than recall of presented items. The effect is persistent across time as well as across repeated opportunities to learn the lists. Results demonstrate the important roles of both encoding and retrieval processes in producing memory illusions (see too Alba & Hasher, 1983; Hasher & Griffin, 1978; Johnson et al., 1993).

Finally, an interesting puzzle arises from examining the results of the current experiments and some prior experiments; the puzzle concerns the relation between false recall and accurate recall in this paradigm. Under certain conditions, false recall levels vary in the same manner as accurate recall levels, whereas under other conditions, false recall levels vary inversely with accurate recall. For example, independent variables enhancing accurate recall sometimes enhance false recall levels, as seen (1) in the effect of prior testing, (2) the benefit of semantic processing (relative to graphemic processing) during encoding, and (3) the benefit of blocked presentation (relative to random presentation). In addition, Schacter et al. (1996) reported a subject variable that had a similar effect: Relative to age-matched controls, amnesic patients showed lower levels of false and accurate recognition. In all four cases, variables affecting the probability of accurate recall affected the probability of false recall in the same direction.

However, under other conditions, independent and subject variables can have opposite effects on accurate and false recognition. In some cases, variables enhancing
accurate recall attenuate false recall, or vice versa. For example, in the multitrial free recall paradigm of Experiment 3, as subjects learned the items in the lists, they also decreased their levels of false recall. Similarly, Norman and Schacter (1996) have shown that relative to young subjects, older subjects showed lower levels of accurate recognition and higher levels of false recognition. Finally, in Experiment 3 the probability of accurate recall dropped from Trial 5 on Day 1 to a later test on Day 2, whereas the probability of false recall rose between the two tests. Perhaps under conditions in which accurate and false recall vary together, the benefit shown in accurate recall might be a result of associative processing. That is, the same processes that enhance veridical recall also drive up false recall. Resolution of when the various patterns are observed must await future research.
Footnotes

1. The one exception to this claim is that *robber* was presented first in the *thief* list, and *steal* was presented second (instead of vice versa as in Roediger and McDermott's, 1995, paper; the purpose of switching the positions was to avoid beginning the list with a homophone.

2. There is one exception: The 13th word in the *foot* list, *smell*, was replaced with *print* in this experiment because it was realized that *smell* is highly associated with *taste*, which occurs in the *sweet* list.
References


Psychology, 2, 331-350.


Hall, J. & Ware, W. (1968). Implicit associative responses and false recognition by young


of Verbal Learning and Verbal Behavior, 6, 175-184.


Journal of Experimental Psychology, 70, 122-129.


Psychological Reports, 25, 803-810.


Reconciling Ballard’s (1913) and Bartlett’s (1932) results. Psychological Science, 3, 240-245.