Theories of Memory
Volume II

edited by

Martin A. Conway
Susan E. Gathercole
&
Cesare Cornoldi

1998

Psychology Press
a member of the Taylor & Francis group
The Role of Associative Processes in Creating False Memories

Henry L. Roediger, III, Kathleen B. McDermott, and Kerry J. Robinson
Washington University in St. Louis, USA

INTRODUCTION: A SURVEY OF ASSOCIATIONISM

Associationism is the doctrine, popular in both ancient and modern times, that mental processes and phenomena can be explained by reference to hypothetical bonds, called associations, that link basic mental representations (Whitlow, 1992). The idea that memory and thinking have their basis in associations dates at least to the time of Aristotle (384–322 BC). Aristotle (1966, p.328) hypothesised that “acts of recollection, as they occur in experience, are due to the fact that one movement [that is, one thought] has by nature another that succeeds it in regular order.” However, associationism as a doctrine did not really emerge until the 1600s, with the rise of the British associationists. Thomas Hobbes (1588–1679) wrote of the “trayne of thoughts” (1651) that was associative in character. John Locke (1632–1704) introduced the phrase “the association of ideas” in his Essay concerning human understanding (1690), and David Hartley even attempted to provide a physiological explanation for associations in 1749. The British associationist tradition culminated in James Mill’s thoroughgoing associative theory in his Analysis of the phenomena of the human mind (1869).

Not long after the publication of Mill’s book, Ebbinghaus (1885/1913) reported the first empirical studies of human memory. His basic assumption was that memory was associative in nature, and he used serial lists of nonsense syllables as his basic tool of study because such lists were learned (he thought) by associating each syllable to the next one and its successors. Associations could be either direct (one item to the next) or remote (between two items that are
separated by one or more items). Through ingenious experiments, Ebbinghaus provided evidence (in patterns of transfer) for the concept of remote associations. The powerful influence of Ebbinghaus's experimental results, superimposed on the backdrop of the British associationists' penetrating philosophical analyses, defined the study of human memory as the study of the formation and retention of associations. Mary Calkins (1863–1930) developed the paired-associates learning technique to study the formation and retention of associations more directly (Calkins, 1894). The interference theory of forgetting, which dominated the study of human learning and memory for many years, was fundamentally associative in nature; the techniques of serial and paired-associate learning were used to study effects of proactive and retroactive interference (e.g. Melton & Irwin, 1940; Postman & Underwood, 1973; Underwood, 1957).

The behaviouristic approaches to learning in animals were also largely associative in character. E.L. Thorndike (1874–1949) assumed that rewards "stamped in" (1903) the connection between stimulus and response. I.P. Pavlov (1849–1936) provided a physiological/associative analysis (1928) of how an originally neutral stimulus could, through its pairing with a stimulus that caused an automatic physiological reaction, acquire the ability to produce a similar reaction. The ascendancy of behaviourism in American psychology in the first half of the twentieth century gave associationistic analyses great force. Many of the hotly debated theories used the concept of association as a core assumption (e.g. Guthrie, 1935).

In the 1950s and 1960s the cognitive revolution began, and for a time it seemed that associationism might be supplanted by other concepts. Psycholinguists argued that associative theories could not explain the complex phenomena of language, in general, and language learning, in particular (e.g. Bever, Fodor, & Garrett, 1968). Moreover, some psychologists studying human memory believed that newer techniques, such as free recall, would resist associative analyses (e.g. Slamecka, 1968; Tulving, 1966, 1968). Despite its vague nature, the concept of schema, popularised by Bartlett (1932), seemed a worthy replacement to some (e.g. Neisser, 1967, Ch. 10). Others preferred Miller's (1956) description of higher-order units in memory (e.g. Tulving, 1964). However, associative theories were not to be denied. Anderson and Bower (1973) provided a tour de force in their book, Human associative memory, showing how associative analyses could be applied to many different memory phenomena. Other models of memory, such Raaijmakers and Shiffrin's (1980) Search of Associative Memory (SAM) model, also built on the assumption that memory is fundamentally associative in character. The rise of connectionism (Rumelhart & McClelland, 1986), with the avowedly associative character of its formulations, also revealed the power of associative analyses.

As this brief survey of some 2300 years of thought reveals, the concept of association has a pre-eminent status in explaining phenomena of learning and memory. Associations are generally viewed, in all these theories, as a powerful
force that benefits memory. Strength of associations determines accuracy of memory. Theoreticians dating from Aristotle speculated on the laws of association, on what factors caused strong associative bonds (similarity of elements, contiguity of elements, contrast among elements, and so on). Powerful associations have uniformly been considered good for learning, retention, and retrieval.

In this chapter we raise the spectre that there is a downside to powerful associations. In particular, we explore the idea that strong associations can lead to false memories. Can powerful associations mislead, as well as lead, memory? Until recently, researchers and theorists working in the domain of associationism have rarely asked this question. We can date interest in the phenomenon to the publication of papers in the late 1950s and early 1960s, but we also note that (at least until after 1965), the issue received very little attention. Some 23 centuries of thought about the nature of associations apparently had not led scholars to realise that powerful associations can be a source of errors, until the introduction of relatively recent experimental results.

Indeed, experimental psychologists studying memory have not, until recently, been much interested in studying errors at all. For example, in the period of 1850–1900, several hundred papers were published on errors of perception in the form of perceptual illusions (Coren & Girgus, 1978), whereas virtually no papers were published on errors of memory (memory illusions). Experimental psychologists have rarely been interested in drawing strong theoretical conclusions from error analyses, although a few exceptions do exist in the literature (e.g. Conrad, 1964; Melton & Irwin, 1940). More generally, psychologists have often considered errors only to reflect processes of “guessing” that might inflate correct responding. Researchers studying recognition and cued recall were frequently concerned with the corrections that should be applied to nominally accurate responses because a portion of responses judged correct are thought to arise by chance from guessing. As a result, in many studies of cued recall (e.g. Roediger, 1973; Tulving & Pearlstone, 1966 among numerous others), errors were not treated as an interesting measure in their own right but only as revealing a process that created interpretive problems for measuring correct responses. In recognition memory the guessing problems and corrections have been deemed even more severe, and entire models of recognition memory have arisen from the theory of signal detection and its utility in correcting for rates of guessing (e.g. Lockhart & Murdock, 1970). In sum, errors were rarely considered to be of interest in memory, unlike in perception research, where their systematic study has stood for well over a hundred years as a crucible for theory construction and testing.

Roediger (1996, p.76) defined memory illusions as "cases in which a rememberer's report of a past event seriously deviates from the event's actual occurrence.” Either the memory for details of the event can be distorted, or (in the most dramatic case) people can remember events that never happened at all. As
we will review later, associative processes can make people believe that an event (the occurrence of a word in a list) actually happened, if the nonoccurring event was merely associated to other events in a sequence. These processes therefore reveal a powerful memory illusion.

As far as we can tell, several papers published in the late 1950s to mid-1960s were the first to raise the possibility that associative processes could lead to erroneous memories. However, the impact of these papers was rather variable. Only one paper, Underwood’s (1965) “False recognition produced by implicit verbal responses”, created much interest at the time. Briefly, if a subject saw a word such as smooth in a long list on which he or she was making recognition judgements for each item, the later appearance of rough was more likely to elicit a false alarm, relative to the case in which the originally exposed word was an unrelated control (such as weak). (We provide a more thorough review of this research later.) Underwood’s (1965) research was replicated and extended, with perhaps a dozen or so papers on the topic appearing in the following decade.

Two other papers appeared before Underwood’s (1965) but did not attract as much attention. Bilodeau, Fox, and Blick (1963) used a cued recall paradigm in which they tried to predict errors in cued recall by analysing the normative strength of association between the cue and a nonpresented (but associated) word. This research is also reviewed in detail later, but the basic question they framed was whether a word strongly associated to a cue word would be intruded in recall even if the word had not appeared in the study list. The answer was yes, and the type of normative association between the two words could be used to predict the pattern of errors.

The third paper addressing associative influences on errors had appeared even earlier. Deese (1959b) presented subjects with lists of 12 words, all of which were associated to one nonpresented word. For a few of the lists, Deese (1959b) observed that on an immediate free recall test, subjects often falsely recalled the critical word that had not been presented—the associate word around which the list had been constructed. Therefore, Deese (1959b) showed how false recall could occur in a free recall paradigm, albeit in only a few lists. Although Deese’s (1959b) report was known to a few researchers, it did not spur much further research at the time, unlike Underwood’s (1965) paper. We speculate on the reasons for this state of affairs later. However, the study of false recall caused by associative factors did not assume much importance in the 1950s through 1970s. For example, in Deese’s (1965) book, The structure of associations in language and thought, he did not mention the findings of his previous paper on intrusions and did not deal with the idea that associative processes could lead to error, despite his extensive treatment in the book of many other topics about associationism in experimental psychology. Similarly, Anderson and Bower’s compendious Human associative memory (1973) did not include the issue of errors produced by associative processes in its 511 pages, citing neither Deese (1959b; although Deese, 1959a, was cited), Bilodeau et al. (1963), nor
Underwood (1965). The issue of how associative processes might cause errors was not of much interest at the time.

The aim of the present chapter is to review what is known about the role of associative processes in causing false recognition and false recall. We delimit our subject matter to include only experiments using lists of associated words. Of course, the issues of false recall and false recognition have been studied extensively within a prose memory tradition dating back to Bartlett (1932). This approach received renewed interest in the 1970s and continues today. We make occasional contact with that literature, but its complete review is beyond the scope of this chapter (see Alba & Hasher, 1983). Similarly, other bodies of research—such as that using Loftus’s (1993) misinformation paradigm—have explored related issues. Again, we draw on this related work when it is germane but primarily review the literature on associative processes in remembering in straightforward laboratory paradigms using word lists.

We organise our chapter in the following manner. First, we consider Underwood’s (1965) false recognition paradigm and the numerous related experiments that followed it. We turn next to the cued recall paradigm introduced by Bilodeau et al. (1963). After a review of these literatures, we turn to Deese’s (1959b) report and discuss how the paradigm was developed into the modern version introduced by Roediger and McDermott (1995). We then review the literature that has developed on false recall (and false recognition) since their report. Although Deese’s (1959b) experiment predates Bilodeau et al.’s (1963), which in turn antedates Underwood’s (1965), the development of these literatures generally proceeded in the opposite chronological order, so we order our review accordingly: false recognition, false cued recall, and finally errors produced in free recall. However, because Roediger and McDermott (1995) adapted the free recall paradigm to measure both false recognition and false free recall, we also consider other measures in this last section. Finally, at the end of the chapter, we briefly consider some of the theories that have been used to explain phenomena of false recall and false recognition and note some strengths and weaknesses of each approach.

**FALSE RECOGNITION PRODUCED BY ASSOCIATIVE PROCESSES**

With the conceptualisation of the implicit associative response (IAR) by Underwood in 1965, the power of natural language associations between words became the subject of increasing research interest. According to Underwood, the perception of a word may result not only in the activation of the representative response (the meaning of the word itself), but also in the unintentional activation of a word that is associatively related to the presented word, the implicit associative response. For example, when the word give is encountered, the meaning of that word is activated as well as the meaning of that word’s strongest
associate; in this case, take. Underwood contended that the IAR was not simply a hypothetical construct but indeed an actual event: the unintentional arousal of a nonpresented word in conscious awareness.

Accordingly, these implicitly occurring associates began to be recognised as a source of interference and error on recognition memory tests, appearing as predictable false alarms to nonstudied items. Underwood (1965) argued that if a certain associate is activated when a study word is encountered, an individual may be likely to confuse the words in memory and believe that the associate had been presented. Evidence of implicit associate activation on a later memory test—in other words, high probabilities of false alarm to words like take in the previous example—could be interpreted as strong support for the IAR theory.

With this idea in mind, Underwood (1965) constructed a 200-word continuous recognition test, designed to measure rates of false alarms produced by IAR activation. In this paradigm, subjects heard a list of words while they indicated whether or not each word had appeared previously in the list. Four types of words occurred in Underwood’s test list (although to the subjects the list was simply a long series of words): stimulus words, critical or experimental words, control words, and filler words. The stimulus words were those like give; they were intended to produce a specific IAR when encountered. Critical words were those believed to be the subject of the IAR, for example, take. Control words were similar to critical words, except that they were not expected to be subject to any IAR activation from the list: They were included to provide a baserate measure of false alarms to unrelated words. Filler words were believed to be neutral and were included and repeated multiple times to increase the feeling of repetition within the list. The data of interest were the subjects’ responses to the critical lures: words that had never been previously presented but that had been preceded in the list by stimulus words expected to elicit them as an IAR. The rate of false alarms to these words was compared to that of the control words, which were not expected to have been implicitly triggered by previously studied items.

Underwood (1965) was also interested in manipulating the nature of the relationship between the stimulus and the critical words, and examining the effect that this might have on false alarm rates to the critical words. Accordingly, he tested four types of stimulus words with respect to their putative IAR (see Table 9.1): antonyms of the critical words, sets of multiple associates converging on the critical words, superordinate words (such that the preceding stimulus words were instances of a category and the critical word was the category name), and sensory impressions (the set of stimulus words all demonstrated a particular characteristic and the critical word was the name of that characteristic, e.g. barrel, doughnut, dome, globe, and spool might elicit round).

It should be noted that in all cases except for the antonyms, multiple stimulus words were expected to elicit a single IAR. For example, in the converging associates case, animal, cat, and bark were all expected to elicit dog as an IAR; in the superordinate condition, maple, oak, elm, and birch were meant to converge
9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES

TABLE 9.1
Stimuli and Selected Results from Underwood (1965)

<table>
<thead>
<tr>
<th>Stimulus Words</th>
<th>Experimental Word</th>
<th>Prob.</th>
<th>False Alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Alarm Baserate:</td>
<td></td>
<td></td>
<td>.13</td>
</tr>
<tr>
<td>Antonyms Presented Once</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bottom</td>
<td>top</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>give</td>
<td>take</td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>day</td>
<td>night</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>man</td>
<td>woman</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>Antonyms Presented Three Times</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rough</td>
<td>smooth</td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>hard</td>
<td>soft</td>
<td>.37</td>
<td></td>
</tr>
<tr>
<td>slow</td>
<td>fast</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>Converging Associates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>butter, crumb</td>
<td>bread</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>bed, dream</td>
<td>sleep</td>
<td>.23</td>
<td></td>
</tr>
<tr>
<td>sugar, bitter, candy</td>
<td>sweet</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>animal, cat, bark</td>
<td>dog</td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>dark, heavy, lamp, match</td>
<td>light</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>warm, chill, freeze, frigid, hot, ice</td>
<td>cold</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>Superordinates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maple, oak, elm, birch</td>
<td>tree</td>
<td>.19</td>
<td></td>
</tr>
<tr>
<td>cotton, wool, silk, rayon</td>
<td>cloth</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td>robin, sparrow, bluejay, canary</td>
<td>bird</td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Sensory Impressions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>barrel, doughnut, dome, globe, spool</td>
<td>round</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>atom, cabin, germ, gnat, village</td>
<td>small</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>bandage, chalk, milk, rice, snow</td>
<td>white</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>.08</td>
<td></td>
</tr>
</tbody>
</table>

on the critical word *tree*. In the case of the antonyms, some stimulus words were presented once, whereas other antonyms (corresponding to different experimental words) were presented three times before the experimental word.

Underwood (1965) reported that the frequency of false alarms to the critical items was highly variable and depended in part on the specific relation between the studied stimulus and the nonstudied critical words. The primary data are summarised in Table 9.1. Few false alarms were observed in the sensory impression condition, leading Underwood to conclude that sensory impression words rarely occur as IARs. He further found that false alarms tended to occur more readily when an antonym to the critical item had been previously presented on three occasions relative to when it had only been presented once, although, as
may be noted from Table 9.1, specific words were confounded across conditions. Underwood also reported that four or five different but converging stimulus associates produced higher rates of false alarm to a common critical item than did the presentation of only two or three converging associates. Based on these patterns, Underwood suggested that the frequency of an IAR’s occurrence, or the number of times it was elicited in the preceding study list, is a strong variable in predicting the likelihood of false alarm errors to the critical items. Again, however, it should be noted that materials were confounded across the conditions, and examination of Underwood’s (1965) data reveals strong inter-item differences. For instance, the single-presentation antonym *day* clearly produced more false recognition responses than did the three-presentation antonym *false*. Despite the irregularities, Underwood’s (1965) method and theoretical conclusions provoked a great deal of interest and closer examination of the phenomenon.

Anisfeld and Knapp (1968) further examined the nature of the relationship between the eliciting stimulus and the critical word. Using a paradigm similar to that of Underwood (1965), they studied the relative abilities of associates and synonyms of the critical word to produce false alarms to that word on the recognition test. Unlike Underwood (1965), Anisfeld and Knapp (1968, Experiment 1) included both the critical word that was an associate (e.g. *white*) to the stimulus word and the critical word that was a synonym (e.g. *dark*) to the same stimulus word (in this case, *black*) in the same test list.

Anisfeld and Knapp (1968) expected that synonyms should produce high rates of false alarms not because they elicited the critical word as an IAR but because they exhibited a considerable amount of meaning-overlap with the critical words. As explained by Anisfeld and Knapp (1968, p.172):

The constant use of paraphrasing in everyday life communication suggests that in coding for memory under normal conditions speakers retain primarily the semantic content of the message. Since synonyms have a large area of meaning in common, they would seem natural candidates for confusion ...

This outlook has a clear similarity to explanations of false recognition in the prose literature (e.g. Bransford & Franks, 1971). The data from Anisfeld and Knapp’s (1968) study and other similar experiments are shown in Table 9.2. As predicted, Anisfeld and Knapp reported significant false recognition for both the associate critical items and the synonymous critical items.¹

In a second experiment, Anisfeld and Knapp (1968) examined more closely the associative relationship between the eliciting stimulus words and the critical false alarms. They compared three types of associate stimulus words: those that tended to elicit the critical item on a free association task (e.g. stimulus: *bitter*; critical word: *sweet*), those that tended to be themselves elicited by the critical item on a free association task (e.g. stimulus: *sweet*; critical word: *bitter*) and
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Underwood, 1965</td>
<td>.13@</td>
<td>.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.27*</td>
</tr>
<tr>
<td>Anisfeld &amp; Knapp, 1968</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>.03#@</td>
<td>.06*#</td>
<td>.09*#</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.04&amp;@</td>
<td>.07*&amp;</td>
<td>.03&amp;</td>
<td>.07*&amp;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fillenbaum, 1969</td>
<td></td>
<td>.17*&amp;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grossman &amp; Eagle, 1970</td>
<td></td>
<td>.17*&amp;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.07&amp;@</td>
<td>.12&amp;</td>
<td>.14*&amp;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hall &amp; Kozloff, 1970</td>
<td>.03</td>
<td></td>
<td>.07*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hall &amp; Kozloff, 1973</td>
<td>.06</td>
<td></td>
<td>.08#</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paul, 1979</td>
<td>.17</td>
<td></td>
<td>.20*</td>
<td></td>
<td></td>
<td></td>
<td>.11*</td>
</tr>
</tbody>
</table>

*Found by the authors to differ significantly from the baseline.
& Preceding (stimulus) word presented 2 times.
# Preceding (stimulus) word presented 3 times.
@ Baseline is a weighted mean of control conditions involving unrelated words.
those that were bidirectional associates or shared both types of relationship (e.g. white–black).

Their motivation in comparing the two types of associates was to determine if the observed false alarms were the result of some encoding of the critical item during study or if they were due to a “reminder” of the studied item occurring during the later test presentation of the critical item. In the former case, false alarms would occur if the preceding item tended to produce the critical item, resulting in it being encoded during study along with the list items. The latter situation would occur if the preceding item tended to be produced by the critical item, inducing a feeling of familiarity when the critical item was encountered for test. In other words, does the phenomenon of false recognition arise more from encoding or from retrieval processes? If false alarms were observed when associates producing the critical item served as the stimulus items, it would suggest that the critical items were encoded during study of the preceding words. On the other hand, if false alarms were observed only when associates produced by the critical item were used on the test, then some mechanism occurring during testing of the critical item would be implicated. In the latter case, Anisfeld and Knapp (1968) argued, the occurrence of the false alarms could simply be a result of testing for them, as, in this case, the nonpresented items should not have been previously encoded.

Anisfeld and Knapp (1968) reported that both associates that produced the critical item during encoding and bidirectional associates (words that both produced and were produced by the critical item) resulted in a higher probability of false alarms to the critical item (.07), whereas associates specifically produced by the critical item did not (.03). Thus, they concluded that some variety of “initial coding” of the critical item during presentation of the preceding item probably does occur.

In discussing the implications of these data, Anisfeld and Knapp (1968, p.178) made no mention of Underwood’s (1965) IAR hypothesis. Instead, they offered a feature-complex hypothesis to explain the findings:

In order to account for these errors it must be assumed that the word is not the ultimate unit of coding. If it were, associates and synonyms should not produce more errors than control words. Rather, our finding ... supports the conception of words as complexes of features.

According to this scheme, the presentation of a word results in the activation of a pattern of features which vary in their salience and significance. Over time, the activation decays, and eventually any item(s) sharing a sufficient number of features with the studied item would be recognised as familiar. Such a conceptualisation would explain the tendency for critical items to be falsely recognised, and as such is an alternative to the IAR hypothesis. Underwood (1969) adopted a similar stance, arguing that memory traces comprised bundles of attributes.
One important difference between the IAR and the feature complex conceptualisations is that the IAR hypothesis predicts that only one specific word will occur as an IAR during encoding, and that only it would be prone to false recognition. According to the feature pattern hypothesis, however, any word that shares a sufficient quantity of features with the studied word could be falsely recognised. Similarly, the feature overlap hypothesis suggests that synonyms—which share many features—should be falsely recognised at very high rates. However, as the data outlined in Table 9.2 show, synonyms do not show particularly high rates of false recognition; in Anisfeld and Knapp’s study (1968, Experiment 1), false recognition of associates was actually somewhat higher than that for synonyms.

Fillenbaum (1969) did report evidence for this degraded feature complex hypothesis. Using a continuous recognition paradigm, he compared the ability of synonyms, antonyms, and control words (other associates) to elicit false alarms when they preceded the critical item on the recognition test. Importantly, the control words in this case were matched with the synonyms and antonyms for associative strength to the critical item. Fillenbaum found that both synonyms and antonyms elicited significantly more false alarms than the associate control words. Assuming that synonyms and antonyms have more features in common with the critical item than do the other associates, and because the synonyms and antonyms were matched with the control words on associative strength, he argued that the false alarms observed in this paradigm were the result of degraded feature complexes rather than simply associative relationships between the stimulus and critical items. However, the increase was relatively modest at 5% (see Table 9.2), and one might expect synonyms (frigid—cold) to share more features than antonyms (hot—cold), but the outcome does not support this prediction. As Fillenbaum (1969) did not include unrelated control items, it is impossible to determine the impact of the purely associative relationship on false alarms in this study. Nonetheless, Fillenbaum conceded that some part of the effect was probably due to associative relationships between the words, although he interpreted this effect as being mediated by features common to the stimulus and critical items rather than due to the occurrence of an IAR or some other strictly associative mechanism.

Independently of Fillenbaum (1969), Grossman and Eagle (1970) also examined the effect of the prior study of synonyms, antonyms, and other words matched for associative strength on the false recognition of a critical word. Using both a continuous recognition paradigm and a recognition task wherein the study and test components were separated by a five-minute interval, they found that the study of both synonyms (.11) and other associates (.14) led to significant levels of false recognition relative to unrelated control words, but that the study of antonyms (.12) did not. They interpreted these results as support for the feature complex hypothesis, arguing that synonyms and other associates have more features in common with a critical item than do antonyms (see Table 9.2). These conclusions clearly contrast with those of Fillenbaum (1969) and Anisfeld and Knapp (1968).
Grossman and Eagle (1970) also attempted to correlate the probability of false recognition of a critical item with the association strength between it and its preceding stimulus item. They found no relationships that even approached significance (but see Deese, 1959b, which they did not cite). On the basis of these data, Grossman and Eagle argued against Underwood’s (1965) IAR hypothesis, which they interpreted as predicting a strong positive relationship between associative strength and the probability of false recognition. Instead, they contended that false recognition occurs when the new (critical) word shares a sufficient quantity of features with the encoded representation of a previously studied stimulus.

The situation was complicated further by Cramer and Eagle’s (1972) report of no false recognition for synonyms or antonyms in a similar paradigm. (Separate data for synonyms and antonyms are not available in their report, but the combined rate of false alarms for both types was .12; the false alarm base rate was also .12.) Additionally, Cramer and Eagle found that critical items that resembled the stimulus items phonemically produced very high rates of false recognition (.21). Methodological differences between this study and the others discussed here may be implicated in producing both the unusually high unrelated base rate and the high rate of false alarms for phonemically similar items; specifically, Cramer and Eagle (1972) used a speeded recognition procedure in which subjects were pressured to respond as rapidly as possible. Thus, Cramer and Eagle (1972) argued that perceptual or phonemic, rather than associative, properties were at work in producing the false alarms observed in this paradigm. Similar findings were reported by Gillund and Shiffrin (1984). Again using the continuous recognition paradigm, Gillund and Shiffrin found that error rates to phonemically and graphemically related items exceeded those of synonyms at both slow and rapid response rates. When subjects were required to respond very quickly (within 900ms), both phonemic and graphemic errors increased substantially (approximately a .09 increase), whereas false alarm rates to synonyms increased only modestly (about .02). However, this interaction between type of response and rate of presentation during the test did not reach significance.

Despite the popularity of the feature complex hypothesis, the importance of whole-word associations continued to be recognised. Hall and Kozloff (1970) attempted to expand on the findings of Underwood (1965), who reported that false recognition increased when a preceding associated item (an antonym) was studied three times rather than just once. In fact, an examination of Table 9.2 reveals that false recognition of related words is very small with only one presentation of the stimulus. Hall and Kozloff (1970) directly manipulated the number of preceding presentations and found that the probability of false recognition peaked with three presentations (.10) and dropped off at a systematic rate when the preceding item had been presented five (.07) or seven (.05) times, as shown in Fig. 9.1. Hall and Kozloff (1970) attributed this somewhat surprising
pattern to the ability of the critical item to "remind" the subject of the stimulus word, especially when the stimulus word had been studied several times.

According to Hall and Kozloff (1970), subjects base their recognition responses on a criterion of perceived situational frequency for each item, such that items that have a perceived situational frequency exceeding some subjective level are judged old, and those that do not are judged new. In this model, perceived situational frequency counts are increased not only by the perception of the word itself, but also by its occurrence as an IAR to an associated item. Hall and Kozloff (1970, p.278) thus proposed that, due to IARs, some nonstudied (critical) words may register frequency counts above the criterion:

We propose that such prior occurrence of experimental [critical] words is a necessary but insufficient condition for their false recognition, because in most such cases the subject is able to avoid the incorrect judgment 'old' through a second comparison. It seems reasonable to assume that IARs occur during recognition just as they do earlier and that the word most likely to occur as an IAR is the … stimulus word which previously elicited the experimental word as an IAR.

![Graph showing probability of false recognition](image)

**FIG. 9.1.** Probability of false recognition of the critical item as a function of the number of previous occurrences of the stimulus word. Adapted from Hall and Kozloff (1970).
In other words, when the critical item is presented for recognition, it is in focal attention as a stimulus word, and could then potentially elicit an IAR of the previously studied original stimulus word, which would remind the subject that it was the stimulus word that had occurred in the list. Hall and Kozloff (1970) went on to suggest that the more times the stimulus word is presented during study, the more likely it is to occur itself as an IAR during recognition. Thus, despite the fact that the critical item may enjoy a relatively high perceived situational frequency as a result of all the presentations of the preceding associate, this level should be much smaller than the situational frequency of the presented associate itself. The direct comparison is presumed to be made during the recognition test when the original stimulus word is brought into awareness by the presentation of the critical item. As a result, the probability of false recognition of a critical item should decrease when the preceding associate has been presented many times. This argument is in line with the earlier findings of Anisfeld and Knapp (1968, Experiment 2), who found that rates of false recognition are very low when the critical item has the tendency to produce the preceding associate as an IAR (see Table 9.2, column 6: Associates Produced by the Critical Word).

Along the same lines, Hall and Kozloff (1970) further argued that the use of a set of converging associates, rather than the multiple presentations of a single associate, would result in high perceived situational frequency counts for the critical item without a correspondingly high activation of any one studied stimulus word. In a second paper, Hall and Kozloff (1973) found support for this hypothesis. Using a continuous recognition paradigm, subjects saw either three or six presentations of a single associate (e.g. stimuli: lamp, lamp, lamp; critical item: light), or one or two repetitions of a set of three converging associates to the critical item (e.g. stimuli: lamp, heavy, dark, critical item: light). As predicted, Hall and Kozloff (1973) found that the probability of false recognition was higher when the set of three converging associates had been studied (.11 for one presentation of the set or .13 for two presentations of the set) than when a single associate had been repeated several times in the preceding list (probability of .08 for three and for six repetitions of the stimulus word; baserate for unrelated words = .06). Hall and Kozloff (1970, 1973) interpreted their results as support for both the IAR hypothesis and a perceived situational frequency criterion model of recognition decisions.

MacLeod and Nelson (1976) provided evidence at odds with both the IAR and the feature overlap explanations of false recognition. In a continuous recognition paradigm, they manipulated the lag (or the number of intervening words) between the preceding associate and the critical item. As shown in Fig. 9.2, MacLeod and Nelson reported that the probability of false recognition was near zero at no delay, peaked at a lag of 5 items, and then gradually decreased when the lag increased to 10 and 15 items. (For more recent evidence making a similar point, see Brainerd, Reyna, & Kneer, 1995.) MacLeod and Nelson (1976) argued
that the nonmonotonic function observed in these data served as evidence against any single-step theories of false recognition, such as the occurrence of an IAR during study or the overlap of features between items. Instead, MacLeod and Nelson (1976) proposed two potential models that could account for the data, one describing the gradual diffusion and decay of feature representations over time, and the other based on differential forgetting rates for specific types of features over time. Several years later, Paul (1979) also examined rates of false recognition as a function of the lag between the stimulus and the critical item. Unlike MacLeod and Nelson (1976), Paul obtained to find a consistent effect of lag on false alarm rates. Paul (1979) concluded that probabilities of false recognition appear to be unaffected by variation in lags of 2 to 120 intervening items.

Vogt and Kimble (1973) addressed the fact that false recognition rates were so low. They proposed that the reason these rates rarely reached 20% was that the associative language structure probably varied highly from subject to subject. In other words, the most common normative associate to a word may not be the strongest associate to that word for a given subject, resulting in a dilution of the false recognition effect when it was based on purely normative relationships. According to the IAR theory, a more powerful way to elicit the false alarm should be to use, as the critical item, the individual person’s preferred strongest associate to the stimulus word.

FIG. 9.2. Probability of false recognition of the critical item as a function of the number of intervening items (lag) between the stimulus and the critical words. Adapted from MacLeod and Nelson (1976).
To test this hypothesis, Vogt and Kimble (1973) developed a two-part experiment designed to tap individual subjects’ idiosyncratic associative hierarchies. In the first part, subjects rank-ordered five possible associates to a set of stimulus words provided by the experimenter. Two months later, the participants returned and each completed a specialised continuous recognition test in either the auditory or the visual modality. The 450-item test list always included the same set of stimulus items, but the specific critical items varied from subject to subject, based on their previous responses to the stimulus items. For each participant, 16 critical items were included that had been ranked first, 16 were included that had been ranked third, and 16 were included that had been ranked fifth. As may be seen in Fig. 9.3, items that were first-order personal associates generated high levels of false alarms (approaching .40); third- and fifth-order associates produced progressively lower levels. Auditory presentation of the test list appeared to increase the number of false alarms relative to visual presentation. Vogt and Kimble (1973) were appropriately troubled by the finding of higher levels of false alarms to control words than to fifth-order associates, but noted that the discrepancy was probably due to the unusual nature and the low frequency of occurrence of many of the fifth-order words. Based on these data, Vogt and Kimble (1973) concluded that close personal associates were highly effective at eliciting false recognition responses, and that relative distances within individual associative hierarchies could be used to predict rates of false recognition.

Several conclusions may be drawn from this section on false recognition. First, the false recognition phenomenon that Underwood (1965) discovered was replicated and extended by many others in the decade or so following the publication of his paper. Second, with a single presentation of the stimulus word, the false recognition effect to a related (but not presented) critical word was often quite small, with an enhanced false alarm rate of only a few per cent. Third, presenting several related words increased the false recognition phenomenon, but in the early experiments researchers used 3–5 associates at the maximum. Fourth, presenting the same stimulus word repeatedly to arouse false recognition produced a paradoxical effect: Two presentations increased later false recognition of the related word, but greater numbers of presentations actually produced a decrease in false recognition (Hall & Kozloff, 1973). Fifth, studies seeking to ask whether synonyms and antonyms produced false recognition rates beyond that of either other associates or unrelated items have produced mixed and inconsistent results. This issue probably bears re-examination. Sixth, two ideas were used to explain the phenomenon of false recognition in the early research: the notion that IARs occurring during encoding created false recognition later, as proposed by Underwood (1965), and the idea that false recognition was due to overlapping sets of features between test items and encoded items as proposed by Anisfeld and Knapp (1968). Much of the early
FIG. 9.3. Probability of false recognition of the critical item as a function of the subjective associative ranking of the critical item in the subject's idiosyncratic associative hierarchy. Adapted from Vogt and Kimble (1973).

research could be described in terms of both frameworks, and definitive evidence to decide the issue conclusively was not provided at the time (and still is not available).

FALSE CUED RECALL

Research on the critical topic of how retrieval cues provide access to forgotten memories is usually dated from the pioneering work of Tulving and Pearlstone (1966), who compared free recall of categorised word lists to recall under conditions when category names were provided as retrieval cues. They showed that when numerous categories were included in the list, subjects remembered many more items under conditions of cued recall than under free recall; this finding formed the basis for their distinction between availability of information in memory (what is stored) and its accessibility (what can be retrieved under a particular set of test conditions). The fact that cued recall exceeded free recall indicated that subjects did not exhaust their knowledge under free recall conditions. But a problem arises: Suppose, when subjects receive the category
name retrieval cues, they simply begin guessing by providing responses that belong to the category. In this way, guessing could inflate category cued recall and therefore overestimate the power of retrieval cues. Tulving and Pearlstone (1966) were fully aware of this possibility and provided safeguards to rule it out. However, another possibility, which they did not consider, is that powerful retrieval cues could lead to false recall. That is, when subjects see the category name *bird* and write down a word that did not occur on the list (e.g. *robin*), they might not be guessing but rather “remembering” a word that did not actually occur on the list. Powerful retrieval cues may elicit false memories as well as veridical ones. This idea still has not been explored in any depth, but some work that appeared before Tulving and Pearlstone’s (1966) experiments suggests that retrieval cues can elicit false recall.

Bilodeau et al. (1963) used normative association data to predict error patterns in a cued recall paradigm. However, in this experiment (and in most of the other experiments reviewed in this section on cued recall), the researchers did not use instructions that emphasised conscious recollection. Rather, subjects were told: “A short while ago you studied some words on page ___. The words below may help you remember those words. Write down the words that appeared on page ___. You must fill in every line. If you cannot remember, write in the first word that the printed word makes you think of” (Bilodeau et al., 1963, p.424). These instructions obviously encouraged free association as well as recollection and would now be considered inclusion instructions within the framework of Jacoby’s (1991) process dissociation procedure. We report findings from these studies as suggestive of processes that might occur in explicit cued recall, but obviously the experiments would need to be repeated with more appropriate instructions, which emphasise remembering, before any firm conclusions can be drawn.

On the basis of the relative associative strengths of a set of words, Bilodeau et al. (1963) classified target (or cue) words into three types. R1-dominant target words have a single strong associate, such as *light* (*dark*, frequency = .57). For these words, the second most common associate has a much lower frequency of occurrence, e.g., *light* (*bulb*, frequency = .06). Bilodeau et al. predicted that words such as *light*, used as cues, would produce large numbers of errors if the correct response were something other than their primary associate. As the Bilodeau et al. paradigm involved using the second most common associates as the correct responses, they expected that these R1-dominant words would produce the highest levels of false cued recall, in the form of intrusions of the primary associate.

The second type of target words, R2-dominant words, had two primary associates that were almost equal in strength (e.g. for *ocean*, *water* was the most common associate, frequency = .31, but was followed closely by *sea*, frequency = .30). Bilodeau et al. (1963) expected this group of words
to produce the most accurate cued recall with the fewest numbers of intrusions. The final class of words were R3-n-dominant words; these targets had many equally strong associates. For words in this third class, the two most common associates occurred at relatively low frequencies, e.g. for *cheese*, the associates with highest frequencies were *cracker* (frequency = .11), and *milk* (frequency = .10), as compared to the sum of the frequencies for all remaining associates of *cheese* (cake + mouse, etc., frequency =.79) (Bilodeau & Blick, 1965). Bilodeau et al. (1963) expected this third class of words to produce large numbers of intrusion errors, consisting of many different associates.

In their basic paradigm, Bilodeau et al. (1963) had subjects study words that were the second most common associates to the set of cue words representing the three classes just described. Then, given the words as cues, subjects were asked to recall as many of the studied words as possible. Based on their classification system, Bilodeau et al. (1963) were able to successfully predict the patterns of errors that would occur on the cued recall task. As predicted, the R1-dominant words produced the highest error rates, specifically high rates of intrusion for their most common associate. R2-dominant words produced the most accurate recall, and R3-n-dominant words produced high error rates with a wider variety of intrusions.

Several subsequent studies further explored this paradigm. Bilodeau and Blick (1965) examined the effects of delay between study of the list and the cued recall test. They found that the original pattern of results reported by Bilodeau et al. (1963)—the highest level of false recall for the R1-dominant words, followed by the R3-n-dominant words, with the lowest levels of false recall occurring for the R2-dominant words—was replicated at delays of 2 minutes, 20 minutes and 2 days. After 28 days, however, Bilodeau and Blick (1965) reported that subjects' cued recall protocols simply began to resemble free-association responses. This is hardly surprising, as the “cued recall” results themselves had a built-in component of free association. As subjects forgot the list words, free association responses dominated subjects’ “recall”.

Several later studies used similar experimental paradigms (Fox, 1968; Bilodeau & Fox, 1968), but with the same instructions that encouraged some mixture of recall and free association. In essence, these researchers seem to have been studying a combination of what today would be called conscious recollection and priming on a conceptual implicit memory test (e.g. Srinivas & Roediger, 1990). Although researchers drew conclusions about false responding in cued recall, we believe that virtually all of these experiments need to be repeated with instructions that warn against guessing, before the conclusions are accepted as demonstrating true errors of memory. The properties of false memories aroused in cued recall procedures still await systematic study.
FALSE RECALL ON FREE RECALL TESTS

The previous sections have dealt with errors induced on recognition and other cued tests. Retrieval cues such as strong associates and the "copy cues" (Tulving, 1976) used on recognition tests are powerful memorial cues. Although most of the literature produced by experimental psychologists documents the power of these cues in eliciting veridical memories, we have seen in the previous sections that these cues can also elicit false memories. This pattern provides further evidence for the claim that the factors that normally operate to improve memory also increase the probability of false memories (Roediger, McDermott, & Goff, 1997).

In this section we review evidence that false memories can be produced via associative processes in free recall of word lists, too. Further, we show that systematic errors can be produced (a) immediately after presentation of the list, and (b) when subjects are strongly discouraged from guessing. The beginnings of this research can be traced to a report by Deese (1959b). Deese was not interested in errors per se but rather in finding evidence for the associative nature of memory. In another paper that received much more attention at the time, and which has been cited much more frequently in the intervening years, Deese (1959a) showed that the degree to which items within a list were associatively related to one another correlated with overall recall of items in the list. Strong associative bonds within a list aided recall. Deese (1959a) also noted that intrusion errors sometimes occurred on recall tests and that there were no theories to explain why.

Deese (1959b) proposed that the associative nature of the list could be used to predict not only accurate recall, but also intrusions in 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 comprised 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 consisted of the 12 most common forward associates to the critical item, the lists differed in their backward association strengths to the critical item. 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 (1959b) 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 administered immediately after presentation of each 12-word list. Deese’s (1959b) predictions were confirmed; Figure 1 in his paper 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 critical 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. Deese’s report of intrusions in free recall did not arouse much interest for the next 25 years, and given the renewed enthusiasm for studies of false recall today, it is worth asking why. Probably most researchers saw the 1959b paper on intrusions as a minor addition to the more important 1959a paper, which showed how associative structure affected recall of list items. As already noted, experimental psychologists have traditionally been much more interested in correct performance, and errors typically have been considered a nuisance rather than an object of serious study. (As noted in the first few pages of the chapter, Deese [1965] himself failed even to cite the 1959b paper in his book on *The structure of associations in language and thought.* ) However, 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 on four critical nonpresented words. Cramer (1965) reported that 51% of her subjects erroneously recalled at least one (of the four possible) critical words, and she did cite Deese (1959b) as a predecessor. However, despite a few desultory citations (e.g. Johnson, Hashtroudi, & Lindsay, 1993), not many researchers knew or cited Deese’s (1959b) paper until the mid-1990s.

**ROEDIGER AND McDERMOTT’S (1995) EXPERIMENTS**

In the Spring of 1993, Endel Tulving mentioned Deese’s (1959b) report to us. We looked it up and found what we considered to be surprising results: Although the levels of false recall were low to moderate for most of the associative study lists Deese had created, the levels for some of the lists were quite high. Deese (1959b) did not specify the instructions subjects had been given; there was no evidence they had been warned strongly against guessing. If subjects had perceived the recall task as a word association task (in which they simply generated associatively related words) more than as a task in which veridical recall was required, then the high intrusion levels observed for some lists might be expected, and so might the strong positive correlation of intrusions to word association norms. We hasten to add that we did not believe that Deese (1959b) had not measured explicit recollection; rather, we were simply puzzled by the
high intrusion rates. Single-trial free recall often reveals very few instances of intrusions (at least for unrelated items), and subjects seem to set a high threshold and do not guess wildly. As Cofer (1973, p.538) put it, “subjects in recall experiments seem reluctant ... to produce material when they are uncertain that it is correct.” (See also Roediger & Payne, 1985.)

If Cofer is correct, then the results from selected lists in Deese’s (1959b) study are even more surprising. Therefore, the aim of our first experiment was to replicate and extend Deese’s (1959b) high intrusion levels with certain lists. We presented six of the associative lists used by Deese to students in a Memory class at Rice University. These lists consisted of the 12 strongest associates to the words chair, mountain, needle, rough, sweet, and sleep. The students were asked to pay close attention to each list and, following presentation of each list, to recall as many words as they could without guessing.

Somewhat to our surprise, we replicated the basic phenomenon reported by Deese: Students erroneously recalled the critical nonstudied item as having been presented 40% of the time, on average. We also examined the output position of the critical nonpresented items and found that when subjects produced the critical items on the recall tests, items tended to occur towards the end of the recall protocols; 63% of the time the items appeared in the last fifth of subjects’ recall. Further, we compared the level of false recall to the level of accurate recall as a function of serial position. Critical items were recalled with a probability equivalent to that of studied items that had been presented in the middle of the lists. This was a surprising finding that has been generally replicated (McDermott, 1996a; Roediger & McDermott, 1995, Experiment 2; Saldaña, McDermott, Pisoni, & Roediger, 1997; Schacter, Verfaellie, & Pradere, 1996c). If we take recall from the middle serial positions as an index of retrieval from long-term store, as in classic two-store models (Atkinson & Shiffrin, 1968; Glanzer, 1972), then the nonpresented items are recalled from long-term store with the same probability as items from the middle of the list.

Following the study and recall of all six lists, the students took a recognition test on which studied items, critical nonpresented items, and unrelated nonstudied items appeared. For each word presented, subjects assigned a rating of sure old (4), probably old (3), probably new (2), or sure new (1). Collapsing across level of confidence, the probability with which subjects correctly endorsed the studied items as being “old” (.86) was no greater than the probability with which they mistakenly classified the critical lures as “old” (.84). Further, over half of the critical nonstudied items (.58) were assigned to the “sure old” category. The recognition results indicated that subjects seem to remember the critical nonpresented associate much as if it had actually been presented.

Encouraged by these results, in which Deese’s false recall findings were replicated and extended to false recognition, we sought to examine further the nature of illusory memories. Among the questions asked in the second experiment were the following: What is the role of initial recall in producing the high
levels of false recognition? (That is, did the preceding recall test cause the high levels of false recognition?) Second, what is the phenomenological experience of subjects when erroneously classifying the critical items as having been studied? (Do they really think they vividly remember the occurrence of the item from the presentation sequence, or does it just seem very familiar?)

To address the first question, we manipulated prior testing history on the recognition test. Subjects studied 16 lists in Experiment 2 but received an immediate free recall test following only half of the lists. Therefore, on the final recognition test, we could compare false recognition for lists previously tested and those not previously tested. On the basis of the testing effect, or the finding that accurate recall (and sometimes recognition) is enhanced by prior testing (Thompson, Wenger, & Bartling, 1978), we hypothesised that testing might solidify false memories, as well (see also Schooler, Foster, & Loftus, 1988).

As shown in Table 9.3 our results supported this hypothesis: Although the probability of false recognition on the lists that had not been followed by an immediate free recall test was substantial (.72), it was not as high as the probability for the lists that had been previously tested (.81). A similar finding was observed for studied items: The hit rate for lists not previously tested (.65) fell short of that for lists previously tested (.79). In addition to the obvious parallels in the effects of testing for studied and critical items, it is worth noting the similarity of magnitude for hits and false alarms in this paradigm: Subjects appear to treat critical lures as list items.

To address the issue of phenomenological experience of subjects, we utilised the remember/know procedure (Gardiner, 1988; Tulving, 1985; see Gardiner & Java, 1993, and Rajaram & Roediger, 1997, for comprehensive reviews). Subjects were given standard old/new recognition instructions but were also told

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Initial Test</th>
<th>Old</th>
<th>Remember</th>
<th>Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studied</td>
<td>Yes</td>
<td>.79</td>
<td>.57</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.65</td>
<td>.41</td>
<td>.24</td>
</tr>
<tr>
<td>Critical</td>
<td>Yes</td>
<td>.81</td>
<td>.58</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.72</td>
<td>.38</td>
<td>.34</td>
</tr>
</tbody>
</table>

Probability of accurate and false recognition (and remember/know judgements) as a function of whether an initial recall test preceded the recognition test.
that for each item classified as old, they should indicate whether they could vividly recollect some specific aspect of the presentation episode of that item (by writing an R, for remember) or whether they simply knew the item had been presented earlier but lacked the ability to recollect the actual presentation (by writing a K, for know). The question of interest was whether subjects would ever claim to remember something specific about the instance of presentation of the critical items that had not been presented. Our hypothesis was that the great majority of critical items erroneously recognised would be classified as known, given that there was no presentation instance to remember. Subjects’ false alarms would be based on some general feeling of familiarity with the items, not a recollective experience (e.g. Jacoby, 1991). Indeed, prior experiments had shown that false alarms are predominantly classified as known (e.g. Gardiner, 1988; Rajaram, 1993). To our surprise, subjects were quite willing to claim that they could remember the presentation of the critical items—they did so on 48% of the lists. Further, subjects gave remember responses to critical nonpresented items with the same probability as they claimed to remember items that actually had been studied (.49). Therefore, unlike in most prior work, we were able to elicit high levels of false remembering.

A final noteworthy finding from this experiment is that the testing effect was localised in remember responses. For both studied and critical items, the probability of know judgements did not change substantially as a function of prior testing (see Table 9.3). Instead, the enhancement seen in overall recognition was restricted to remember judgements. Despite the fact that subjects were specifically told to make their remember/known judgements with respect to the study episode (i.e. not to report whether they remembered writing the item on the previous free recall test), taking the test made subjects more likely to think they were recollecting the original study episode.

In summary, Roediger and McDermott (1995) expanded the work of Deese (1959b) and changed the basic paradigm he introduced into one that is becoming widely used to induce and assess false memories in recall and recognition. In general, we will refer to experiments in which related words are presented to elicit false recall and false recognition (often with metamemory judgements) as the Deese-Roediger-McDermott paradigm, or the DRM (pronounced DREAM) paradigm for short.3 Obviously, the Roediger and McDermott (1995) experiments owed a debt to Deese (1959b), but the paradigm we introduced and which is in common use now is quite different from the one Deese used. Among other differences, it includes only lists intended to produce high levels of false recall, followed by recognition and metamemory judgements.

Next we review the studies that have explored false recall and false recognition in variants of the Deese-Roediger-McDermott paradigm. Most of the studies to date have manipulated various independent variables to examine their effects on the dependent variables mentioned earlier: probability of false recall,
output position in false recall, probability of false recognition, confidence, and probability of remember responses. A few studies, however, have introduced new dependent variables, usually measuring metamemorial aspects of performance. First we review the various other metamemory judgements that have been examined, then we turn to a discussion of the many independent variables that have been manipulated, followed by discussion of the subject variables examined. Finally, we conclude this review section with a discussion of the few studies that have attempted to identify the neural correlates of illusory memory by using this paradigm. Although reviewing the literature in terms of classes of variables may not be as exciting as doing so in terms of theoretical implications, we justify our organisation by noting that at this early stage of inquiry it will be more profitable to try to gain an understanding of the basic phenomena to be explained. Later, after we have reviewed the evidence, we turn to its implications for some of the most popular theories of false memory.

ADDITIONAL METAMEMORY JUDGEMENTS IN FALSE RECALL AND FALSE RECOGNITION

In an attempt to explore further the enigmatic finding of high proportions of remember judgements to critical lures, Norman and Schacter (1997) asked subjects to justify their recognition judgements (remember, know, or new). Even when pressed to describe the aspect of the item’s presentation that was remembered, subjects still frequently assigned remember classifications to the critical nonpresented items. However, the information provided by subjects in this condition was usually not specific to the individual item (e.g. the way the item sounded when spoken), but often referred to thoughts or associations accompanying the word (e.g. for the word music: “This word occurred in the list with note and piano; I remember thinking about the music I listened to this morning”).

In a second experiment, Norman and Schacter (1997) attempted to quantify subjects’ recollections of the nonpresented items by giving them a variant of the Memory Characteristics Questionnaire (MCQ) developed by Johnson, Foley, Suengas, and Raye (1988). Subjects were asked to specify for each item remembered whether they could recollect various sensory characteristics (e.g. sound) or spatiotemporal context (e.g. list position) of the items. Specifically, subjects were asked to rate (on a 7-point scale) the degree to which they remembered (1) the sound of the word; (2) the position the word occupied in the study list; (3) the word that had come immediately before or after the target word; (4) any reaction they might have had when studying the word; (5) a specific thought occurring during study of the word; or (6) associating this word with other studied words. Three variables (sound, list position, and specific thought) discriminated between studied and critical items in that subjects reported
retrieving more such information for studied items than for critical items. Nonetheless, subjects did claim to be able to ascribe substantial detail to the critical nonpresented items.

Mather, Henkel, and Johnson (1997) obtained results that converge nicely on Norman and Schacter's (1997) basic findings. In their study, subjects were asked to rate (on a 5-point scale) the degree to which they recalled (1) the sound of the word; (2) feelings or reactions encountered when previously hearing the word; (3) any associations made when hearing the word; and (4) rehearsing the word. As in Norman and Schacter's (1997) data, both the sound of the word and feelings when hearing the word were given higher ratings for studied items than critical items. Also consistent with Norman and Schacter's (1997) data, the association rating did not discriminate studied from nonstudied items. The final variable, rehearsal at study, also did not discriminate between studied and nonstudied items. This last finding is perhaps one of the most interesting in this study because it suggests that the locus of the false memory effect may not simply be an implicit activation of the critical item at study but rather that subjects are consciously, effortfully processing and rehearsing these nonpresented items during the study phase. We will return to this idea in a discussion of the possible mechanisms for the effect, later in the chapter.

The primary conclusion from the MCQ studies is that when subjects are asked about the specific perceptual features of test items, they can recall more such attributes for items that were heard in the DRM lists than for those often recalled but never presented. Still, they frequently make errors in recall and recognition by endorsing critical nonpresented items as having occurred in the list, and they also claim to recollect attributes of the items' presentations, albeit not at the level of studied items. It should be pointed out that the differences observed between studied and critical items on these measures are based on statistical aggregates: they are not substantial enough to allow for the classification of a memory as veridical or false on an item-by-item basis.

A final metamemory judgement that has received attention is that of voice attributions. In a variant of the basic DRM procedure, Payne, Elie, Blackwell, and Neuschatz (1996a) presented a long, 64-word list, constructed around six critical items. Items were always blocked according to list, but the voice in which the items occurred was manipulated. Subjects were shown a videotape of two people ("Jason" and "Carol") speaking the words in the lists. Subjects were then given three successive recall tests (an aspect of the experiment to be discussed later, in the repeated testing section). Following the third recall test, subjects were asked to assign voices to the words they had recalled (i.e. whether Jason or Carol had spoken the word during study), with the option of leaving blank any words to which they could not assign a voice. Consistent with the finding that subjects think they remember critical nonpresented items, they also claim to recollect the voice in which the items were presented: In Payne et al.'s experiment, subjects assigned a voice to 87% of the critical items they had recalled. The comparable
rate for studied items was 94%. Although the probability for studied items was greater than for critical nonstudied words, subjects still showed a remarkable tendency to report the voice in which critical nonpresented words had been presented.

Payne et al. (1996a) had several different ways of presenting the lists; in one condition, all of the words in a given sublist (e.g. all words related to sleep) were presented by a single speaker. This condition allowed an examination of the likelihood that subjects would classify the critical nonpresented items as having been spoken by the person who spoke the words that corresponded to that critical item. Surprisingly, subjects classified the nonpresented items this way only 53% of the time. In contrast, the probability of correctly classifying the studied items to the appropriate speaker was 84% in this condition. These findings are something of a puzzle.

Using a slightly different procedure, Mather et al. (1997) found a much higher attribution figure on a recognition test. In their study, 76% of subjects claimed that the critical nonpresented item had been spoken by the speaker who had spoken the associated words. In Mather et al.’s experiment, however, subjects were not given the option of claiming not to remember the speaker of the item: They were told on the recognition test to choose whether the item had been presented by the female speaker, the male speaker, or not at all. Thus, the figures are not directly comparable. It would also be interesting to know in Payne et al.’s experiment how the remaining 47% of items were classified—were they attributed to the other speaker, or were most classified in the “don’t know” category? Payne et al. did not present these data.

Regardless of the specific patterns of response classifications, the primary conclusion from these two studies is that subjects are quite willing to attribute a voice to the presentation of the critical nonpresented items (see also Saldaña et al., 1997). The extent to which these judgements are influenced by the voices that spoke the associated studied items is still something of an open question: Mather et al. (1997) conclude that it is a highly influential factor; Payne et al.’s (1996a) results, however, seem to suggest otherwise. Further work is clearly needed before any strong conclusions can be drawn.

In summary, several studies have examined a number of dependent variables in the DRM paradigm (or close variations of it). These variables include the probability of recall, probability of recognition, output position, and probability of remember responses. In addition, voice attributions and specific questions about the perceptual and cognitive aspects of the initial study phase have been examined. The majority of the results converge on the claim that although it is sometimes possible to distinguish between studied and critical items, the false memories created in this paradigm are extremely robust (often to the extent of being indistinguishable from true memories). In the following section, we review how important independent variables that have powerful effects on veridical memory influence false memories.
FACTORS AFFECTING FALSE RECALL AND FALSE RECOGNITION IN THE DRM PARADIGM

We begin this section with a discussion of the variables that have been examined in the encoding phase, followed by a review of the experiments with manipulations of retention interval, and conclude with factors in the test phase and how they influence false recall and false recognition in this paradigm. This organisation of factors in terms of encoding, retention, and test phases corresponds to standard experimental practice, but of course it must be borne in mind that manipulations of factors at a particular stage (e.g. retrieval) are not independent of earlier stages of processing (e.g. encoding). Still, for our purposes, this scheme provides a useful way to organise the literature.

Study Variables

Included in this section are variables that have been manipulated before and during study presentation. These variables include manipulation of instructions during the study phase and manipulations of the study material itself. Researchers have asked many fundamental questions about false recall by manipulating basic study variables within the DRM paradigm. We consider these studies here.

Instructional Manipulations at Study

An interesting but as yet little-examined question is whether these false memories are immune to a warning, or knowledge of the false memory phenomenon. Gallo, Roberts, and Seamon (1997) examined false recognition on a final test (given after presentation of a series of associative lists) as a function of the study instructions given to subjects. In the uninformed and cautious groups, subjects were given standard study instructions, similar to those used by Roediger and McDermott (1995); in the cautious group, however, test instructions also included clauses asking subjects to attempt to minimise false alarms and informing them that the test would include nonstudied associates of presented items. In a third condition, the forewarned condition, subjects were informed before the study phase of the specific memory illusion under investigation and were given a sample study list and corresponding test to demonstrate the effect. Subjects were told to attempt to minimise their false alarms to these critical nonpresented items on the subsequent recognition test.

Gallo et al. (1997) found that a vague warning to be cautious (given just prior to the test) reduced the hit rate but did not lead to a statistically reliable decrease in the critical false alarm rate, relative to the standard instructions. Numerically, the decrease was present, however: Critical false alarms occurred with a probability of .81 in the uninformed condition and .74 in the cautious condition, as can be seen in Fig. 9.4. (The instruction to be cautious did reliably attenuate the
9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES

proportion of remember responses assigned to the critical items, in comparison to
the uninformed group.) When subjects were explicitly forewarned, they were
able to decrease substantially the proportion of critical lures erroneously
recognised. False recognition was still substantial, however, even in the
forewarned condition, in which subjects classified the critical lures as old .46 of
the time. The primary conclusion from this experiment is that knowledge about
the false recognition phenomenon is sufficient to reduce but not to eliminate the
effect. A nonspecific warning to attempt to minimise false alarms did not reliably
attenuate the false alarm rate of the critical nonpresented items, although strong
conclusions about this effect cannot be made, in part because of the moderate
(.07) yet statistically insignificant effect.

McDermott and Roediger (1997) have obtained data that point to a similar
conclusion. In these experiments, the critical item around which the list was
constructed was presented in half of the lists and not in the other half. All subjects
were informed that the lists contained words that were highly associated to one
critical item and that this item sometimes was present in the list but sometimes
was not. Subjects were told to pay close attention to the lists and to make sure that
they did not remember the critical item as having occurred when in fact it did not.
We also gave our subjects a sample list, and explained to them that people often

![Graph](image)

**FIG. 9.4.** Probability of accurate and false recognition as a function of instructional condition. Adapted from Gallo et al. (1997).
mistakenly recalled the critical item (in our example, *king*) when in fact it had not been presented and to be sure that they did not commit this error.

In our first experiment, 12 lists were presented, followed by an old/new recognition test. Although subjects were better than chance at determining the prior history of the items, they still classified nonstudied critical items as "old" 61% of the time. (When the item had been presented, the hit rate was 78%.) In both this experiment and in Gallo et al.'s (1997) experiment, one factor contributing to the false recognition effect may be in the delay between study and test; all lists were presented prior to the recognition test, so there was a moderate delay between study and test (due to the fact that all lists were presented, followed by one long recognition test). Therefore, in Experiment 2, McDermott and Roediger moved the test so that it occurred immediately following each list. That is, subjects were given a one-item recognition probe immediately following each 15-word study list. Here, too, subjects were better than chance at discriminating whether or not the item had been present in the list; nevertheless, they were far from perfect at this task (with .38 of the nonpresented items erroneously classified as old and .81 of the presented items correctly classified as old). By comparing results of McDermott and Roediger's (1997) two experiments, we can conclude that changing the task to a more direct source-monitoring task (by introducing an extremely short retention interval) made it easier for subjects to determine whether the item had been present in the list, although it did not improve their ability to correctly determine when the item had been present in the list.

Perhaps the most interesting aspect of all these data is the simple fact that despite an explanation of the phenomenon, an explicit warning not to make such errors, and an immediate test, false recognition of the critical lures was still robust. In sum, the warning studies existing so far show that although subjects may be able to reduce false alarms in this paradigm, they are by no means completely successful in doing so. Further work is needed to determine the extent to which warnings can influence study and test strategies so that subjects can most effectively reduce false recall and recognition.

A second instructional variable that has been manipulated at encoding is level of processing (Craik & Lockhart, 1972). Level of processing manipulations are usually instantiated by instructing subjects to think about some meaning-based attributes of the studied words (e.g. their pleasantness)—in the deep condition—or to think about an orthographic or phonemic attribute (e.g. the number of vowels in the word)—in the shallow condition. Relative to shallow processing, deep processing greatly enhances recall and recognition of studied items (Craik & Tulving, 1975). Read (1996) examined the proportion of recalled and remember responses (in the remember/know paradigm given after the recall test) as a function of study orientation for one of the lists most likely to induce false recall, the *sleep* list. Read (1996) postulated that even if there was no difference in level of recall of the critical nonpresented item as a function of level of
processing, it seemed reasonable to expect differences in assignment of remember judgements to the critical intrusions. This hypothesis was based on Gardiner, Gawlik, and Richardson-Klavehn’s (1994) results with accurate recognition, which show that elaborative rehearsal enhances remember judgements (and leaves know judgements unaffected) and that maintenance rehearsal enhances know judgements (but not remember judgements). Therefore, Read (1996) hypothesised that elaborative rehearsal should lead to enhanced recall and remember judgements for the critical nonpresented items, as well as for studied items, relative to the more shallow processing task of maintenance rehearsal. His results, however, failed to support the prediction: Although level of processing affected recall of studied items in the predicted direction, sleep was erroneously recalled with similar probabilities in the maintenance (.76) and elaborative (.73) conditions, and the remember responses were unaffected as well. However, Read also did not replicate Gardiner et al.’s basic finding of increasing remember responses for studied items, so it seems possible that there was not enough power in his experiment to detect differences.

Tussing and Greene (in press) examined the effect of level of processing on false recognition following the presentation of six 12-word associative lists. They found no differences in the probability of false alarms to the critical nonpresented words in their three encoding conditions: pleasantness rating, letter counting, and judging whether each word began with a vowel. However, caution is warranted in interpreting this null effect because the deep encoding condition (i.e. pleasantness rating) did not produce a higher hit rate than the two shallow encoding conditions.

Two studies have demonstrated level-of-processing effects on retrieval of the critical nonpresented items. Toglia, Goodwin, Lyon, and Neuschatz (1995a) found that deep processing (pleasantness judgements) enhanced recall of critical intrusions (as well as studied items) relative to the shallow condition (determining whether the word contains an “a”). Similarly, Thapar, McDermott, and Fong (1997) obtained large effects of level of processing such that judging the pleasantness of the list words produced higher levels of false (and accurate) recall than did counting the number of vowels in the words or determining their colour. This pattern was achieved on a final free recall test given immediately after study of the lists, and also after one- and seven-day delays.

The results of level-of-processing manipulations are inconsistent and puzzling. One possible explanation for the lack of effects reported by some researchers is that regardless of the overt study task, subjects naturally engage in a deep level of processing when presented with blocks of associatively related words. This interpretation would account for the lack of level-of-processing effects on studied items obtained in some of the experiments. We expect that the level of processing engaged in at study will be shown to affect studied and critical items similarly, as Toglia et al. (1995a) and Thapar et al. (1997) have found. Further experiments will be needed to provide conclusive evidence, however.
Study Material

We turn now to characteristics of the study material itself. Among the variables examined thus far are: number of associates comprising the study lists, rate of presentation of the studied items, list structure when multiple lists are presented together (i.e. whether blocking associates together results in differing rates of false recall compared to interspersing items from various lists), effects of repeating presentation of the study words, and dividing attention during the study phase. Finally, we consider whether some associative sets induce higher probabilities of false recall and false recognition than others and possible implications of such a finding. We consider first the very basic question: How does the number of associates in a list affect false recall?

Number of Associates. Deese (1959b) found that the level of false recall in his experiment was well predicted by the mean associative strength of the list. That is, false recall following his 12-word lists correlated highly ($r = .87$) with the mean associative strength linking the list words to the critical nonpresented items. In their first experiment, Robinson and Roediger (1997) introduced a straightforward variation of list length. The lists were either 3, 6, 9, 12, or 15 items long. (Lists always contained the highest associates first; thus the 12-word lists were constructed by dropping the last three items from the 15-word list, etc.) Robinson and Roediger (1997) argued that on the basis of Deese's hypothesis, one might predict that shorter lists would produce higher probabilities of false recall because they have higher mean associative strengths. However, their results showed the opposite: the longer the list, the higher the probability of false recall of the critical nonstudied item (see Fig. 9.5). In a second experiment, list length was held constant by adding unrelated words to the number of associates in the lists from Experiment 1. In this way, the average associative strength of the lists to the critical nonstudied item was much lower than in Experiment 1; if mean associative strength determines false recall, then false recall should have been much lower in Experiment 2 than Experiment 1. However, by comparing results of the two experiments, Robinson and Roediger were able to determine that the filler words added to equate list length (in Experiment 2) did not attenuate the intrusion levels relative to Experiment 1, although recall of the presented associates was affected by this manipulation (see Fig. 9.5). This finding suggests that it is the total associative strength of the list, not the mean associative strength, that most accurately predicts false recall.

A second interesting finding to emerge from Robinson and Roediger's (1997) experiments is that the critical nonpresented item does not behave as though it were another (presented) item in the list, as Roediger and McDermott (1995) seemed to find. As the list length increased, the probability of recall of any individual studied item decreased, whereas the probability of false recall increased. The probabilities of false and accurate recall were therefore inversely related.
**Presentation Rate.** Another basic study manipulation is the rate of presentation of the study items. Predictions for the effects of presentation rate on false recall are difficult to make. If one believes that deeper processing promotes higher false recall, it might follow that slower presentation, which allows for more elaborative and relational processing, might enhance false recall. Conversely, it seems intuitive that the faster the items go by, the more difficult it might be to effectively encode each individual item; subjects might be left with a general feeling of the overall theme of the list but little retention of the specific items. The empirical evidence for presentation rate exists in the form of several unpublished studies, at least as of this writing (May 1997). Schwartz (1996, personal communication) presented items at either a 1-, 5-, or 10-second rate and found false recall was inversely proportional to presentation rate: the longer the study period, the lower the probability of false recall. Specifically, the level of false recall fell from .51 to .44 to .26 with the 1-, 5-, and 10-second rates, respectively. Toglia and Neuschatz (1996) also found differences in false recall as a function of study time. Specifically, they found a .72 false recall probability for a fast, 1-second rate, but that probability was sharply diminished (to .49) when the study rate was extended to 4-seconds. Therefore, the current data suggest that in general, the faster the rate of presentation, the higher the rate of false recall. Whether this claim will survive manipulations at the extreme is still undetermined. Robinson, Balota, and Roediger (1997) tested subjects at several
rates under 100ms. Of course, veridical recall is quite poor at such rates, but the interesting finding was that (relative to probability of accurate recall), the probability of false recall remained quite high. Subjects recall very little, but what they do recall is as likely to be the critical nonpresented item as any of the items that were presented. This research is still being conducted with recognition procedures, so further comment is not warranted at this time.

**Blocked/Random Presentation.** In a variation of the basic paradigm discussed here, McDermott (1996a) combined several lists to make one long list, in which the items converged around three critical nonpresented words. That is, the 45-word study list contained 15 items related to each of three critical items (e.g. cold, sleep, and needle). This type of list construction permits an examination of whether and how order of presentation of the studied words affects the probability of false recall. Specifically, to what extent does false recall differ as a function of whether items are blocked according to list or are randomly intermixed throughout the study phase? Opposing predictions can be made. One could argue that random presentation would lead to higher false recall because the overall list structure is more confusing, leaving subjects trying to organise the list subjectively but unsure of the exact items that were studied. The fact that quicker presentation enhances false recall would seem to bolster this prediction. Conversely, it could be expected that blocked presentation would lead to enhanced false recall because it would enhance relational processing and magnify the possibility of implicit associative responses occurring. For example, the probability that point would elicit needle as an IAR would be greater if point were presented in the context of other needle-related words (e.g. thread, pin, sewing ...) than if it occurred in proximity to other words unrelated to needle.

As can be seen in Fig. 9.6, McDermott (1996a) found that blocked presentation of the list led to higher false recall than random presentation. Indeed, in the blocked condition, after a single presentation of the list, the probability of false recall (.57) exceeded the probability of veridical recall (.38). Toglia et al. (1995b) have obtained this blocked/random effect in recall, and Mather et al. (1997) and Tussing and Greene (in press) have found the same pattern in recognition memory.

**Repetition.** Tussing and Greene (in press) examined whether repetition of the study list affects the level of false recognition. On the basis of IAR theory, one might expect that the probability of false recall and false recognition would increase with repetition of the list because the probability of an implicit associative response of the critical nonpresented item would be enhanced with multiple presentations of the study words. However, when Tussing and Greene (in press) presented words from six associative sets (randomly ordered), they
found no reliable difference between false recognition when the list items were presented once (.74) or three times (.67) (see also Shiffrin, Huber, & Marinelli, 1995).

McDermott (1995) presented three associatively related 15-word lists blocked together. As expected, repetition enhanced accurate recall (from .38 for once-presented lists to .61 for lists presented five times); however, repetition decreased the probability of false recall (.22 for repeated lists, compared to .57 for once-presented lists). Explanations of this finding are not obvious, but it is reminiscent of the reduction in false recognition shown by Hall and Kozloff (1970; see Fig. 9.1). However, before we stretch too far to find explanations, it would be desirable to determine if the effect is replicable.

**Dividing Attention.** A fifth study variable that has received interest is the effect of dividing attention during the study phase on later false recognition. This question is interesting because it speaks to the issue of the role of elaborative, relational processing in producing the phenomenon. If organisational strategies during the study phase are contributing to the effect (as would be suggested by the outcomes of the level of processing, study time, and blocked/random experiments), we would expect that dividing attention during the study phase would diminish false recall and recognition. In two experiments, Payne, Lampinen, and Cordero (1996b) found that dividing attention during the study phase does indeed attenuate later false recognition. This outcome can be
considered as supporting Underwood’s (1965) notion that IARs during study elicit false recognition, that IARs would be less likely to occur if attention is divided.

**Characteristics of the Study Lists.** An interesting aspect of Deese’s (1959b) original results, and one that is somewhat overlooked today, is that relatively few of his lists produced high intrusion rates. For example, the famous *butterfly* list (the 12 most common associates to the word *butterfly*) failed to elicit any intrusions in Deese’s sample of 50 subjects. Roediger and McDermott (1995, Experiment 1) used six of Deese’s lists, the ones that had produced the highest intrusion rates in his study. For our second experiment, we created 18 more lists and published all 24 in the Appendix of our paper; these lists have been used extensively in later research, both by ourselves and by others. We have since created 12 more lists, as have other groups of researchers; it is also clear that standard categorised lists of the Battig-Montague (1969) type can elicit high false recognition of high-frequency items from the category norms when these are omitted from the study list (Hintzman, 1988; Shiffrin et al., 1995).

In most of the research reported in this chapter, a set of materials is used (e.g. the 24 Roediger-McDermott lists), and the data are aggregated over materials. However, from even a casual glance at the results, it is apparent that the materials vary dramatically in their effectiveness in producing false recall and false recognition. Some lists produce high levels of erroneous remembering and others practically none. Deese (1959b) showed that in his set of materials, the average tendency of the items in the list to arouse the critical item as an associate predicted intrusions in recall. As of this writing, we know little more than this. Our main point in including this topic is to draw attention to it and to how critical it is to understanding the phenomenon of false memories aroused by associative processes.

The central question is why some lists produce high levels of false recall and false recognition, whereas others do not. After all, the lists are constructed in the same way. What characteristics of lists and/or target words elicit the phenomenon? Can the “good” lists be altered in some way to reduce levels of false remembering? Can the same be done with the “poor” lists? Such knowledge would bring the phenomenon under experimental control. Do abstract materials or concrete materials produce the stronger effect? Can subjects (or experimenters) predict which lists will produce high levels of false recall and false recognition?

We know the answer to none of these questions at this point, although each is currently stimulating experimental inquiry. Stadler, Roediger, and McDermott (1997) have made a start by providing data that will be critical to answering these questions. Stadler et al. performed a norming study in which two large groups of subjects studied and recalled 18 of the 36 lists in the expanded Roediger-McDermott set (the 24 lists in their Appendix and 12 more developed by
McDermott). After all lists had been studied and recalled, subjects received a recognition test over the whole set. Therefore, we have norms on the effectiveness of each list in producing false recall and false recognition. (The recognition results are possibly contaminated by the prior recall test, although other evidence makes us think that this is probably not a critical problem; see Robinson & Roediger, 1997, Table 1). These norms are a starting place for asking questions about the effectiveness of individual lists in producing false memories. Stadler et al. found that false recall and false recognition are quite reliable across subgroups of subjects (with split-half correlations of around .80). However, probabilities of correct recall and recognition correlated only weakly with the probabilities of false recall and recognition (.20 and .15, respectively). False recall and false recognition were highly correlated across the 36 lists, $r = .76$.

Although the Roediger-McDermott lists were created to produce high levels of false recall and false recognition, and in the aggregate they do, the variation among lists is substantial. For example, lists constructed around the words *window, sleep, smell,* and *doctor* all produced false recall at levels at or above 60%. Conversely, six lists constructed the same way, and with the same face validity in producing false recall, led to levels of false recall at or below 25%. The list associated with *king* produced only 10% false recall. The same general pattern held for recognition. Although overall the lists produce high levels of false recognition, the variability in materials is great: It ranged from 84% false recognition (for the *window, smell,* and *cold* lists) to relatively low levels of false recognition (for the *lion, 33%,* and *king, 27%,* lists). Again, the reasons for these wide differences are not apparent from a simple examination of the materials. We also hasten to add that even the materials producing "low" levels of false recall and false recognition do so with higher probabilities than one would find with unrelated materials. The systematic study of variables underlying the differential effectiveness of materials is of critical importance to understanding these phenomena, but the work on this issue is just starting.

**Retention Interval**

We turn now to look at the next stage, the retention interval, and the role it plays in erroneous remembering. We begin by considering a short delay, designed to eliminate the contribution of primary memory, and then turn to the studies that have examined the effects of longer delays (of the order of hours and days) in order to flesh out the time course of false recall and relate it to forgetting curves for studied items.

McDermott (1996a) asked whether a short, 30-second delay would affect the level of false recall. It is well documented that for studied items, such a delay eliminates the recency effect but leaves recall levels for the beginning and middle portions of the list unaffected (Glanzer & Cunitz, 1966; Postman & Phillips,
1965). On the basis of the finding that the proportion of critical items recalled approximates the proportion of studied items recalled from the middle of the list on an immediate free recall test, and the assumption that recall from this part of the list represents recall from long-term store, we might hypothesise that recall of the critical items represents recall from long-term store. If so, we might expect that a short delay would not affect critical item recall (because it has no effect on long-term store but only short-term store).

This prediction was borne out in McDermott’s (1996a) data. As can be seen in the first two data columns of Table 9.4, recall of the critical nonpresented item in the delayed condition (.46) did not differ from recall in the immediate condition (.44). (As expected, the delay eliminated the recency effect but left recall of the pre-recency items unaffected.) Zechmeister and Zechmeister (1996) reported a similar experiment, in which a delay of 90 seconds was used. In their experiment, the short delay affected studied items in the predictable manner (i.e. in the recency portion of the serial position function); in addition, they found that false recall increase across the short delay (from .55 on the immediate test to .70 in the delayed condition). This finding parallels studies that have examined more substantial delays, which have shown that false recall and recognition tend to increase or remain stable, whereas forgetting occurs for the studied items.

In McDermott’s (1996a) experiment, in which subjects took an initial test either immediately or after a short delay, all subjects returned two days later for a final free recall test covering all the studied lists (those lists not initially tested, as well as those tested in the immediate and 30-second delayed conditions). Although levels of recall differed on the final test as a function of initial test condition (a testing effect, which will be discussed in the following section), in all three conditions the probability of false recall exceeded that of veridical recall after this two-day delay. Collapsing across the three conditions, the probability of

| No Test | — | — | .04 | .12 |

TABLE 9.4
McDermott’s (1996a) Results

| Type of Recall Test
<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>
</tbody>
</table>

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.
false recall was .20 and the probability of accurate recall .12. In a second experiment, McDermott compared recall on the final free recall test on the day of study presentation to recall on the same test given one day later. Although studied items were forgotten across the retention interval, there was a small but statistically significant rise in the proportion of critical items recalled.

In a similar experiment, Payne et al. (1996a) observed no change in false recognition over a 24-hour delay (although correct recognition decreased). It is not yet clear when false memories in this paradigm increase with delay and when they remain stable, but the primary conclusion to be taken from these studies is that false recall is extremely robust, either remaining constant or increasing over a time interval, during which memory for the studied events is diminishing. As a consequence, in either case (stability or increase in absolute proportions), there is an increase in false memories relative to veridical memories. These findings are consistent with research in the prose recall literature, in which false recall and recognition of schema-consistent prose passages increase over delays ranging from one week to a year (Barclay & Wellman, 1986; Spiro, 1980; Sulin & Dooling, 1974).

**Variables Manipulated at Test**

**Testing Effects.** We turn now to test variables pertaining to false recall and false recognition. The aspect of the test phase that has been most closely examined is the testing effect, or the finding that performance on a test is enhanced if an initial test is taken relative to when there is no initial test. Roediger and McDermott (1995) reported that this effect occurs not only for studied items, which is well known, but also for critical nonpresented items (see Table 9.3). Although some researchers have replicated this effect (e.g Saldaña, et al., 1997; Zechmeister & Zechmeister, 1996), others have not consistently replicated the testing effect for critical nonpresented items (e.g. Norman & Schacter, 1997; Payne et al., 1996a, Experiment 2; Schacter et al., 1996c). Similarly, Roediger and McDermott’s (1995) finding of a testing effect in remember responses has enjoyed limited replicability (e.g. Payne et al., 1996a obtained it, but Norman & Schacter, 1997, failed to replicate it).

In an effort to make sense of the discrepant findings in the literature, McDermott (1996b) performed a meta-analysis of all experiments examining testing effects on false recognition in the Roediger-McDermott paradigm. Collapsing across studied and critical items, in 17 of the 22 relevant comparisons, there was a numerical advantage for the condition in which an initial test had occurred. When broken down by item type, the mean probability of accurate recognition was .75 and .67 for lists previously tested and not tested, respectively; comparable probabilities for the critical items were .73 and .70. When subjected to Wilcoxon signed-ranks tests, these data show a reliable overall effect for studied items but not for critical items.
A similar analysis was performed to determine the generality of the testing effect on remember judgements. The mean probabilities of remember judgements for studied items were .55 and .44 for tested and nontested conditions, respectively; the comparable levels for critical items were .52 and .46. A Wilcoxon test showed a reliable testing effect for studied items; however, the test for critical items fell just short of significance.

In the aggregate, the results with respect to accurate recognition and remember responses support the findings of Roediger and McDermott (1995) in that prior testing enhances later recognition and recollection (as manifested in remember judgements) of the studied items; the effect on critical nonpresented items is less conclusive.

Why is the testing effect manifested in some experiments but not others? One possibility is that recognition is a relatively insensitive measure with respect to testing effects. Support for this hypothesis lies in the finding initially reported by Darley and Murdock (1971) that under conditions in which little or no testing effect is observed for a final recognition test, a robust testing effect is obtained on a final free recall test. Lockhart (1975; see also Jones & Roediger, 1995) later showed that when the testing effect is obtained in recognition, it is generally restricted to an enhancement of the last few items of the study list. The finding that a recall test enhances later recall but not later recognition is broadly consistent with transfer appropriate processing approaches to memory (e.g. Morris, Bransford, & Franks, 1977; Roediger, 1990): Practice in recalling information transfers well to recalling it again, but does not transfer as well to later recognition.

Given that testing effects are ubiquitous in free recall and inconsistent in recognition, we might expect testing effects for false memory to behave similarly (i.e. to be more easily obtained on free recall tests). McDermott (1996a) has obtained robust testing effects on final free recall. As shown in Table 9.4, on a final free recall test given after a 48-hour delay, the probability of false recall corresponding to a given study list was twice as great (.24) when a free recall test had followed the study presentation than when maths problems had been worked following the presentation (.12). In another series of experiments, McDermott (1996c) has shown that the number of initial tests taken (0, 1, or 3) strongly influences the later probability (on a final free recall test) of both accurate and false recall. In addition, the number of tests affected remember judgements (both accurate and erroneous) in the same way.

On the basis of these results with free recall, we might ascribe the fleeting nature of the testing effect on recognition to the relative insensitivity of the test. It is our opinion that the testing effect is a real phenomenon in the domain of false memory (see also Roediger et al., 1997 for more extensive discussion of this point).
9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES

**Hypermnesia.** A couple of studies have examined whether there is a false memory analogue of hypermnesia. If people are given several tests in immediate succession, will the probability of false recall increase across the tests, similar to the increase often found in the probability of accurate recall (Erdelyi & Becker, 1974; see Payne, 1987, for a review)? Payne et al. (1996a, Experiment 2) addressed this question by presenting subjects with 60 words, which were grouped according to six associative sets (of 10 items per set). Although hypermnesia was not obtained for studied items, there was a reliable increase in critical items produced across the three tests (.27, .28, and .35 on Tests 1, 2, and 3, respectively). Payne et al.'s results also suggested that this increase was not attributable to a relaxation of criterion across tests. In a follow-up experiment with a slightly different procedure, they obtained similar data in that recall of the critical nonpresented items increased across successive tests (.27, .30, and .33 across Tests 1 to 3); this time, they also obtained increases across tests for studied items, or hypermnesia.

McDermott (1996c) also attempted to find an analogue of hypermnesia in false recall. Unlike Payne et al. (1996a), however, McDermott obtained no reliable hypermnesia for critical nonpresented items. The discrepancy between her findings and those of Payne et al. is likely due at least in part to the nature of the study lists: McDermott gave successive immediate free recall tests following the presentation of each 15-word associative set, whereas Payne and his colleagues grouped the associative sets together, followed by a series of final tests. Because hypermnesia (in the standard sense, with studied items) is usually only obtained following long study lists (Roediger, Payne, Gillespie, & Lean, 1982), it is plausible that list length is the distinguishing factor between the two studies. Note, however, the point mentioned earlier that the number of initial tests taken in McDermott's experiment did influence later false recall and remember judgements. Therefore, despite no increase across the successive initial tests, the act of taking the tests did influence later memory for the original event (by enhancing later accurate and false recall).

**Multitrial Learning.** The experiments reviewed here have examined the role of repeated testing in the absence of a subsequent study presentation. One can also employ a multitrial learning procedure to ask whether subjects will be able to use successive study–test episodes to edit out their intrusions in recall. McDermott (1996a) showed that across five study–test episodes, when accurate recall was improving steadily, false recall dropped substantially (from .57 to .32 when the items were blocked according to associative set and from .30 to .20 when the items were randomly ordered). Remarkably, though, even after five attempts at learning the list, subjects still produced robust levels of intrusions. Subjects were able to reduce but not eliminate false recall. This finding is similar to the results of Gallo et al. (1997) and McDermott and Roediger (1997) in showing that the memory illusion is quite resistant to correction.
Dividing Attention. In a final test variable, in which multiple tests were not used, Payne et al. (1996b) examined the effects of dividing attention during the test phase on recognition levels of the critical nonpresented items. Recall that these researchers showed that dividing attention during study attenuated false recognition. In the same experiment, a similar finding was obtained for the test phase. When subjects were required to monitor for the presentation of three consecutive odd digits (presented over headphones, during the visual recognition test), both accurate and false recognition levels were diminished.

These results are interesting in part because they differ from those of investigators exploring other illusions of memory. For example, Jacoby, Woloshyn, and Kelley (1989b) found that dividing attention at test reduced accurate recognition of items but did not affect the illusory memories they were studying (the false fame effect from studying nonfamous names). Therefore, they argued that the false fame effect arises from automatic processes. Applying the same logic to Payne et al.'s experiment, the illusory memories produced in the DRM paradigm would be attributed to conscious recollection because dividing attention diminished the false recognition levels. This conclusion is in agreement with the finding that recognition of the illusory items is accompanied predominantly by remember judgements.

In summary, the studies that have examined the effects of test variables have clearly shown that just as testing influences what is later accurately remembered, testing can affect the likelihood of producing a false memory as well. In addition, dividing attention has detrimental effects on memory for studied items and has similar effects on false recall of the critical nonpresented items.

Subject Variables. Researchers are just beginning to explore how subject variables might affect false recall in the DRM paradigm. There are two basic questions that can be asked: How do subject populations differ in their susceptibility to (and patterns of) memory illusions? Within a subject population, are there any factors that are likely to predict susceptibility to the illusion across individuals? We begin this section by reviewing the brief literature on subject populations, and then consider the question of individual differences within the normal population of subjects.

Norman and Schacter (1997) examined false recall and false recognition in the DRM paradigm in older adults, noting that older adults are more susceptible to false recognition in Underwood's (1965) continuous recognition paradigm (using both synonyms and rhymes; see Rankin & Kausler, 1979), in Jacoby et al.'s (1989a) false fame paradigm (Bartlett, Strater, & Fulton, 1991), and in Loftus's (1993) misleading information paradigm (Cohen & Faulkner, 1989). On the basis of these findings, they hypothesised that older adults would similarly show higher rates of false recall and recognition after studying lists of associated words. Results confirmed this prediction. As shown in the top half of Table 9.5, age and item type interacted such that younger adults recalled more studied items
(.68) than older adults (.51) but recalled fewer critical nonpresented items (.36) than the older subjects (.49). (These data were obtained by collapsing across their two experiments.) The recognition data generally showed the same pattern, with younger adults manifesting higher hit rates and lower critical false alarm rates than the older adults, although the false alarm rate difference between subject groups was evident in only one of the two experiments.

In a similar study, Tun, Wingfield, and Rosen (1995) compared older and younger adults with respect to their probabilities of false recall in the DRM paradigm. Although older adults recalled fewer studied items than younger subjects on immediate free recall tests, the subject groups did not differ in their levels of false recall. It is presently unclear why Norman and Schacter obtained age differences, whereas Tun and her colleagues found none. On the basis of work in other paradigms (discussed earlier) and existing theories of the effects of ageing on frontal control systems in memory (e.g. Moscovitch, 1989, 1995) we suspect that further work will reveal that older adults will show inflated false recall relative to younger adults (in this paradigm, as well as others); however, more research in this area is clearly needed before definitive conclusions can be drawn.

The other subject population that has been examined at this point is amnesic subjects, who show quite a different pattern from the normal older adults. Schacter et al. (1996c) compared normal older adults to twelve amnesic subjects, six of whom exhibited Korsakoff's syndrome, and six of whom had varying etiologies (but all had damage to medial temporal and/or diencephalic structures). The recall data are shown in Table 9.5: Amnesics recalled fewer studied items (.27) than did controls (.52) as well as (nonsignificantly) fewer critical items (.29) than did controls (.33). Schacter et al. note, however, that interpretation of the nonstudied difference is complicated by differences between the groups in other false alarms. An examination of the final recognition test data supports the conclusion that amnesics' memory for the critical nonpresented items (.16), like their memory for the studied items (.16), was lower than older controls (.66 and .57 for studied and critical nonpresented items, respectively). In some sense, the amnesic patients do not have sufficient capability for conscious

<table>
<thead>
<tr>
<th>Reference</th>
<th>Subject Population</th>
<th>Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Studied</td>
</tr>
<tr>
<td>Norman &amp; Schacter (1997)</td>
<td>Young</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>.51</td>
</tr>
<tr>
<td>Schacter et al. (1996c)</td>
<td>Old</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>Amnesics</td>
<td>.27</td>
</tr>
</tbody>
</table>
recollecion to remember the cITICAL nonpresented items! These results lend
support to the idea, recurring in this chapter, that manY of the processes that
support accurate memory seern to overlap with those that support association-
induced false memories. It will be interesting to see how other memory-impaired
subject populations (e.g. patients with dementia of the Alzheimer's type) will
behave in this paradigm. Schacter, Norman, and Kousaal (in press) review what
is known of the neuropsychology of false memories in the DRM paradigm, as
well as in other paradigms.

We now turn to the idea that individual differences within the population of
normal individuals might predict susceptibility to this memory illusion. Hyman
and Billings (1998) explored a number of different measures and showed that in
a different paradigm, the single characteristic that predicted false recall was
performance on the Dissociative Experiences Scale (DES; Bernstein & Putnam,
1986). This scale contains questions such as "Some people have the experience
of finding themselves in a place and having no idea how they got there. How
often does this happen to you?" People who report that such incidents happen to
them often tended to be more likely to produce false recall in Hyman and
Billings' (1998) report. On the basis of this research, Winograd, Peluso, and
Glover (1996) asked whether the DES (or some other measure) might correlate
with false recall and recognition in the Roediger-McDermott paradigm. Results
of their study show that the DES can reliably predict associative false
recognition, $r = .34$.

Research into individual differences and subject variables is clearly just
beginning; there is still much to be learned in this domain. We consider now the
studies that have used the DRM paradigm to explore the neural correlates that
might underlie false memories.

Neural Correlates

Schacter et al. (1996b) used positron emission tomography (PET) to explore the
neuroanatomical substrates that underlie false recognition. On the basis of the
finding (discussed earlier) that patients with medial temporal damage exhibit low
levels of false recall, coupled with reports implicating this region in conscious
recollection (Nyberg et al., 1996; Schacter et al., 1996a), Schacter et al. (1996b)
hypothesised that false recognition, like accurate recognition, would involve (in
part) medial temporal structures. This prediction was borne out in the data:
Accurate and false recall were accompanied by similar patterns of hippocampal
activation.

Of primary interest in this study were possible differences in regions
underlying accurate and false recognition. On the basis of the finding that
accurate memories sometimes contain richer sensory and perceptual detail than
do false memories (Mather et al., 1997, discussed earlier in the chapter; Schacter
et al., 1996c), Schacter et al. (1996b) predicted that auditory regions might be
differentially active for studied and critical nonstudied items. (The lists in this study were presented auditorily and tested visually.) Consistent with this prediction, Schacter et al. found reliable increases in blood flow during recognition for the studied items in left temporoparietal cortex, an area that is generally thought to underlie phonological/auditory processing. No such activation occurred for the critical nonpresented items. Thus, this experiment was able to provide evidence that the neural bases of true and false memories, although similar, may be distinguishable in some instances.

A similar conclusion was reached by Johnson et al. (1997); using event-related potentials (ERPs), they found differences between studied and critical nonpresented items on a recognition test (however, see Düzel et al., 1997, who used a similar approach and failed to find any differences). Johnson et al. (1997) reported greater bilateral prefrontal activity for “old” responses given to studied items than for the same responses given to the critical nonpresented items. However, this result was obtained only when test items were blocked according to item type (old, critical nonpresented, or unrelated new items). When the test items were randomly intermixed (as is usually done in behavioural studies, and was done by Düzel et al., 1997), no differences were found between studied items and critical nonpresented lures. Johnson et al. suggested that researchers should be sensitive to the processing changes that accompany such changes in task design (see also McDermott et al., 1997).

Johnson et al. (1997) also attempted to explore Schacter et al.’s (1996b) finding of temporoparietal differences between old and critical items. Although the waveforms for hits and critical false alarms did not differ from each other, they did differ from correct rejection of the critical lures. On the basis of this finding, Johnson et al. suggested that the blood flow differences obtained by Schacter et al. in their comparison of studied and critical nonpresented items might have been driven by correct rejection of the critical items. This type of argument is clearly speculative but suggests that there is a great deal of work to be done before blood flow differences averaged across many old and new items can be confidently interpreted as definitive markers in distinguishing between true and false memories.

**SUMMARY OF THE EMPIRICAL RESULTS: WHAT DO WE KNOW?**

In the preceding pages we have reviewed evidence from a large number of studies from the past 30 years on the topic of false recall and false recognition produced by associative processes. In this section we summarise some of the primary findings that any theory will need to explain. Following this summary, we briefly review several of the main theories introduced for explaining false recall and false recognition, and note strengths and weaknesses that exist for each theory, based on the foregoing review of results.
1. False remembering induced by associative processes can be found in free recall, cued recall, and recognition. We suspect that every task measuring explicit memory will reveal systematic memory illusions, even when subjects are given strict instructions not to guess.

2. As the number of words associated to a target word increases in a study list, the probabilities of false recall and false recognition increase in direct proportion, at least up to 15 associates.

3. Blocking materials at study (placing associated words together in sublists) increases false recall and false recognition relative to randomly ordered presentation; this finding parallels the blocked/random effect for veridical recall.

4. Increasing the rate of presentation has opposite effects on false recall and accurate recall, at least through ranges of around 1 second to 10 seconds. Faster rates lead to lower levels of veridical recall but higher levels of false recall.

5. Increasing retention intervals leads to forgetting of studied material but often has little effect on erroneous recall. False memories either remain stable over relatively short intervals, or actually increase. “Forgetting” of false memories may obey different laws from normal forgetting of studied material.

6. False memories aroused by associative processes often have the feel of real (veridical) memories. Subjects report remembering the moment of occurrence of words that were not presented, the voice presenting the words, and other characteristics of the event. Subjects’ phenomenological experiences are often quite similar to those for real memories, although in some cases differences are detectable.

7. Testing effects occur for false recall as well as for veridical recall: Recall of a nonstudied word on a first test increases the probability it will be recalled on a later test. The same is probably true for the effect of recall on false recognition, but the data for recognition are less consistent.

8. These false memories are also quite robust, durable, and resistant to correction.

9. Whereas older adults show poorer recall of studied events relative to younger adults, false recall does not show similar drops with age and may even increase. Conversely, amnesic patients with damage to medial temporal structures show poorer accurate recall and less tendency towards associative memory illusions. Within the normal range of subjects, performance on the Dissociative Experiences Scale correlates with the tendency to produce illusory memories.

Of course, many other results have been reviewed besides the nine listed here, but these nine seem relatively secure. Other findings, such as whether antonyms or synonyms produce false recognition effects similar to those produced by associates, or whether increasing level of processing has an effect on false recall or false recognition, are in conflict in the literature, and we must await further research for their resolution.
ASSOCIATIVE PROCESSES IN FALSE MEMORIES 233

THEORETICAL IMPLICATIONS

Associative processes can induce people to remember words that were not presented in a list studied only seconds earlier. Given the traditional emphasis on how associations produce excellent retention, this finding may seem surprising and counter-intuitive. However, many different theories have been extended to explain errors produced by associative processes. We cannot review all of these theories here in detail; instead, we will confine our remarks to a brief evaluation of the main strengths and weaknesses of several approaches to explaining the facts listed earlier. These approaches often have been advanced independently of one another but are not mutually exclusive.

The Implicit Associative Response Hypothesis

Underwood (1965) argued that encoding of a word arouses to conscious awareness an associate of that word, the implicit associative response. This simple idea can go far in helping to account for many of the facts just listed. The facts that increasing list length and the blocking of items at presentation increase false memories in variations of the DRM paradigm can be handled naturally by assuming that more IARs are aroused during blocked presentation and when more items are studied. In addition, the realistic feel of these false memories can all be accounted for by assuming that IARs are aroused during encoding. The item seems to have occurred because the subject recently thought of it. On the other hand, one might expect that slowing the rate of presentation would increase (not decrease) the occurrence of IARs; however, false remembering increases with faster rates, contrary to this prediction. Finally, if the critical item occurred during encoding as an IAR, it might be expected to behave like other items in the list as a function of independent variables. However, in some cases it does not. For example, the falsely recalled item is not forgotten according to the same function as studied items. Although the implicit arousal of the falsely remembered item must, almost by definition, play some role in this phenomenon, the idea by itself seems too simple to account for all the reported effects. The IAR hypothesis must be complemented by other theories.

Automatic Spreading Activation Theory

Underwood (1965) assumed that implicit associative responses were consciously produced during encoding. However, another approach to explaining phenomena of false recall and false recognition is to assume that activation spreads throughout a large semantic network automatically and unconsciously, as in models such as those of Anderson and Bower (1973) and Collins and Loftus (1974). The “node” of an item such as sleep could be primed by having 15 related words recently presented, and this activation might trigger false recall and false
recognition. Such a theory provides a natural account for such findings as the effect of list length on false recall and false recognition, and for the fact that blocked presentation of the list leads to greater false remembering than does random presentation. This theory is testable in that one could correlate amount of priming in tasks such as lexical decision or naming from these lists (on the target words, such as *sleep*) with the probabilities of false recall and false recognition. Do lists that produce considerable priming also produce high levels of false recall and false recognition? Individual lists could also be manipulated to produce greater or lesser priming, and false recall and false recognition could be examined to see if corresponding effects are observed.

One drawback to this theory, and some others discussed later, is that the simple concept of activation does not permit linkages to subjective experience. Subjects in the DRM paradigm have the strong sense of remembering the illusory items, but sheer activation would probably be expected to give rise to experiences of knowing, not remembering. However, if one combined Underwood’s (1965) idea of the IAR—the conscious, if implicit, production of the response—with the concept of automatic spreading activation, then the combined theory may be capable of accounting for most of the results reviewed. For example, the fact that divided attention at study reduces false recognition in the DRM paradigm would be difficult to explain using just automatic spreading activation (why should dividing attention affect an automatic process?), but the hybrid theory could do so by postulating that part of the false recognition was due to conscious processes and part to automatic processes. Such a theory would resemble two process accounts of priming, with controlled and automatic components (e.g. Neely, 1977; Posner & Snyder, 1975). However, details of such an account as applied to false recall and false recognition remain to be worked out.

**Feature Matching Theory**

In the continuous recognition paradigm, Anisfeld and Knapp (1968) argued that matching of features between the test item and studied items caused false recognition. This idea has been systematically elaborated in some global memory models of recognition memory (see Clark & Gronlund, 1996, for a review). The assumption is that matching of meaning features is critical—if the meaning of the test item matches that of the studied items to a sufficient degree, subjects classify the item as having been studied. If we assume that items associated to a critical nonpresented item share many semantic features with the critical item, then when the critical lure is provided on the test, it will often trigger a false alarm from matching of features. This provides the basic account of false alarms.

Although we believe that global matching theories could account for most of the results listed earlier, we note several weak links. First, one might expect synonyms to produce high rates of false recognition relative to other related
words (such as associates or antonyms). Although Fillenbaum (1969) reported 5% greater false recognition from synonyms than associates, the false recognition rate was still rather low (given that meaning features overlapped quite closely), and this effect was not obtained in other work (e.g. Anisfeld & Knapp, 1968, Experiment 1).

A second problem is that global matching models assume that a test item is encoded and produces some familiarity value that varies on a dimension of strength. Strong feelings of familiarity lead to judgements of “old” if the criterion is surpassed. Critical lures in the DRM paradigm would receive very high familiarity values and therefore produce high levels of false alarms. However, the sticking point is the rich phenomenological experience of subjects in reporting false memories. The memories do not merely seem quite familiar in some global sense, but subjects claim to remember considerable detail about the events. Know responses are usually thought to indicate general familiarity, but in the DRM paradigm, subjects claim to remember the nonpresented words. There may be solutions to this issue within global matching models, but in general the mapping from constructs in the models to subjective experiences of the rememberer has not been worked out. The same seems true of connectionist theories applied to false memory phenomena (McClelland, 1995).

Third, Brainerd et al. (1995) reported that under some conditions there is a reversal of false recognition, wherein presenting a related item makes false alarms to a critical lure less likely. Some conditions have been considered earlier in this review, such as several presentations of the related item before the critical lure (Hall & Kozloff, 1973) or presenting the related items immediately prior to the critical lure (MacLeod & Nelson, 1976). As Brainerd et al. (1995) point out, it is difficult to reconcile this finding with global matching models.

**Fuzzy Trace Theory**

Fuzzy trace theory can explain the false recognition reversal just described and has other features that make it an attractive account of false recall and false recognition, which has led several groups of researchers to adopt the theory to explain their results (Payne et al., 1996a; Schacter, et al. 1996c). However, a similar sort of difficulty exists for Reyna and Brainerd’s fuzzy trace theory (see Reyna & Brainerd, 1995, for a review) as for global matching models. According to fuzzy trace theory, experiences are assumed to leave two traces: verbatim traces (or those with specific features) and gist traces (or ones that capture the meaning of events but lack perceptual detail). False memories such as those created in the DRM paradigm are attributed to remembering the gist of experiences. The theory can account for many phenomena, but again founders (in our opinion) on the subjective experience reported by subjects. Remember responses are given by subjects when they believe details of the original experience are available, whereas know responses arise from general feelings of
familiarity. The most natural mapping in fuzzy trace theory is that remember responses should reflect specific traces and know responses should arise from gist traces. Yet false memories in the DRM paradigm are predominantly experienced as remembered rather than known. Further developments of fuzzy trace theory may explain how this pattern of results is possible, but would seem to do so only at risk of abandoning some of the main assumptions about the two classes of memory traces.

Source Monitoring Framework

Johnson's source monitoring framework is directly relevant to the issue of how false memories are created (e.g. Johnson & Raye, 1981; Johnson et al., 1993). From this viewpoint, when subjects report remembering a word that did not appear in a list, they may be experiencing source confusion. This explanation is similar to that invoked by IAR theory: Subjects may internally generate the critical word during study of the list and later confuse the private events of thought with the external presentation of words. Many of the findings listed earlier could be interpreted within the source monitoring framework. One advantage of the source monitoring framework is the great emphasis on the experience of the rememberer and how different encoding operations may affect source confusions. In addition, considerable attention is devoted within the theory to decision criteria whereby the rememberer decides if something actually happened or was only imagined.

In explaining false remembering arising from associative processes, the source monitoring account shares considerable territory with the implicit associative responses hypothesis. That is, in both cases the subjects' confusion is thought to arise from their failure to distinguish whether the event in question (say, the occurrence of sleep within the list of associated words) actually happened in the external world or was aroused internally. As such, the weaknesses attributed to IAR theory also must be dealt with here, viz., the findings in which the event that was assumed to be activated does not behave like other members of the list. For example, the implicit response is not forgotten over time according to the same function as other members of the list. Still, no theory handles that finding gracefully, and the source monitoring framework is attractive because it is able to explain many other false memory phenomena, such as those that arise from creating mental images (e.g. Garry, Manning, Loftus, & Sherman, 1996; Goff & Roediger, in press).

Attributional Analysis of Remembering

Jacoby's attribution theory of remembering (e.g. Jacoby, Kelley, & Dywan, 1989a) is also relevant to the issues at hand, and he has incorporated memory illusions, such as the false fame effect (Jacoby et al., 1989b) into his theory.
9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES

Inspired by attribution theories that arose in social psychology, Jacoby and his colleagues have argued that activation arising from memory can be attributed (and misattributed) to various sources. For example, in an experiment reported by Jacoby, Allan, Collins, and Larwill (1988), subjects heard words in a first phase of the experiment. In a later, ostensibly unrelated phase they listened to words through noise and tried to identify them. Words that had been heard in the first phase were identified more easily in the second phase than words that had not been heard. However, subjects attributed this ease of perception to the noise having been less loud on those trials involving repeated words. In this case, an effect of memory was misattributed to variations in perceptual conditions at testing.

In the case of false memories, the situation is typically reversed. Some event comes easily to mind (either from the environment or from some other source) while subjects are engaged in retrieving events from memory from some particular time and place from the past. Because they are engaged in remembering specific events, and because some idea comes to consciousness during this attempt, the subjects attribute the activated information to “having a real memory” from that time and place. In fact, however, the events are not being remembered—or at least it is not from the time and place from which the subject is retrieving; instead, the memories arise from some different source.

To explain the phenomenon of false recall and false recognition of associatively related material, one need only assume that during the course of being tested on a list, the associated word (e.g. sleep) comes easily to mind, just as do list words. Because subjects are actively engaged in remembering and an event comes easily to mind, the item is misattributed to having been in the study list and is therefore “remembered”. Again, this account probably could be used to explain most of the findings described earlier. One potential difficulty for the approach (although not the rendition just outlined) is that many memory illusions adumbrated by Jacoby and his colleagues have been hypothesised to arise from automatic processes in memory (i.e. processes that would be expected to give rise to know judgements rather than remember judgements in Tulving’s [1985] and Gardiner’s [1988] remember/know paradigm). However, by extensions such as the one just outlined, the theory would seem capable of explaining illusions of memory that arise from consciously controlled processes, or remembering. If one assumes that when people are engaged in an attempt to remember, the events that come easily to mind will be attributed to past experiences and will have the experiential feel of past experiences; in this way, false remembering would be explained.

Again, as with the source monitoring framework, Jacoby’s attributional ideas are appealing because they can potentially account for many memory illusions. Indeed, the source monitoring framework and the attributional analysis of memory share many features in common. In our opinion, these theories provide perhaps the most compelling means of explaining how false memories arise, both
from associative processes (as in the paradigms under discussion here) and from other processes (such as interference effects), too.

CONCLUSION

The concept of association is central to an understanding of human memory. The contribution of the present chapter is to review evidence that strong associations, long known to facilitate remembering, can create memory illusions as well. The literature reviewed suggests that a non-occurring event that is strongly associated to other, experienced events will frequently be recalled and recognised as having occurred. Research on this issue enjoyed a flurry of activity in the 1960s and early 1970s, remained relatively dormant for a time, and now has returned to the fore as researchers become engaged in exploring the processes involved. We suspect that this development is a healthy one for the field. As noted, errors have often been considered a nuisance in the study of memory and thought to be useful only for correcting “hit rates” in recall or recognition by factoring out the “false alarms” assumed to arise through guessing. We suggest that this approach is no longer profitable, if ever it was. Rather, we should consider the study of sensing and perceiving as a model for studying remembering (Roediger, 1996). Psychologists interested in those topics have studied perceptual illusions for 150 years and realised that an understanding of veridical perception demands a parallel explanation of illusions. Perceptual illusions have been a fertile ground for developing and testing theories of perceiving; the time is ripe for the systematic study of illusions of memory, as well.

NOTES

1The base rate for recognition errors for words unrelated to the associate stimulus words was .04; the base rate for words unrelated to the synonym stimulus words was .02. Anisfeld and Knapp (1968) determined significance levels for associate and synonym critical words based on these values. The unrelated base rate value in Table 9.2 is a weighted average.

2Grossman and Eagle (1970) used separate base rate measures for each type of stimulus word. The base rate for base for items unrelated to the antonym stimulus words was .08; the base rate for items unrelated to the synonym stimulus words was .05; the base rate for items unrelated to the associate stimulus words was .08. Again, the base rate value in Table 9.2 is a weighted average.

3We thankEndel Tulving for suggesting the nomenclature of calling this the DRM (or DREAM) paradigm.

4The .57 and .38 values are taken from McDermott’s (1996a) Experiment 2 and not from McDermott (1995). However, the comparison is a valid one because the study conditions and the population from which subjects were drawn did not differ for the two experiments; the only difference was the number of list presentations preceding the recall test.
ACKNOWLEDGEMENTS

We thank Lyn Goff, Steve Kanne, Kate Pfeifer, Mitch Sommers, and Jason Watson for helpful comments on earlier drafts of the chapter.

REFERENCES


9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES


9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES


...


9. ASSOCIATIVE PROCESSES IN FALSE MEMORIES


