ACADEMIC
PRESS

Journal of Memory and Language 47 (2002) 469–497

Journal of
Memory and
Language

www.academicpress.com

Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring

David A. Gallo[*] and Henry L. Roediger, III

Department of Psychology, Washington University, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899, USA

Received 17 September 2001; revision received 1 December 2001

Abstract

Associative lists created by the same means are remarkably different in their propensity to elicit false memories in the DRM (Deese, 1959; Roediger & McDermott, 1995) paradigm. We confirmed this variability in Experiment 1 by constructing lists in the typical fashion but with words that were weakly associated to their critical words. Low levels of false recall occurred. In Experiment 2 these results were replicated at three presentation rates (.5, 1, and 3 s per word). Also, slower presentation rates yielded lower false recall for both strong and weak lists. Experiment 3 showed that false recognition rates also varied across lists, as did subjective ratings accompanying false recognition. We interpret these findings as supporting an activation/monitoring framework. Lists vary in a principled way in their tendency to activate the critical item, and slowing the presentation rate permits greater accrual of item-specific information that makes monitoring of retrieval more accurate. © 2002 Elsevier Science (USA). All rights reserved.

Keywords: False recall; False recognition; Presentation care; Associative strength; Subjective judgments

Roediger and McDermott (1995), adapting procedures used by Deese (1959), introduced a paradigm that is now widely used to study one type of false memories. Subjects heard either 12 or 15-item lists and were asked to recall each list immediately after it was presented. With minor exceptions, the lists were all composed of the highest associates of one word in the Russell and Jenkins (1954) word association norms. For example, the items for one list (*bed*, *rest*, *awake*, etc.) were associated to the word *sleep*. In immediate free recall tests, with instructions not to guess, these critical nonstudied words were recalled with a high probability (.40 with 12-item lists and .55 with 15-item lists) and these rates rivaled the probability of recall of items from the middle serial positions of the list. False recall—recall of events that did not actually occur—was as great or even greater than recall of some events that did occur. The same general outcome occurred on recognition tests that were given after many lists had been presented, with false alarm rates to critical items equaling the hit rates for the list items. False recognition of unrelated lures was quite low, so subjects were not simply responding positively to all test items.

The levels of false recall and false recognition reported by Roediger and McDermott (1995)

[*] Corresponding author. Fax: +1-314-935-7588.
E-mail address: dgallo@artsci.wustl.edu (D.A. Gallo).

PII: S0749-596X(02)00013-X

seemed surprising, even to them, because the experiments incorporated features that are usually believed to discourage false memories. Subjects were presented word lists (thought to encourage a reproductive, rather than reconstructive mode of recollection; Bartlett, 1932). Their recall tests immediately followed each list and included instructions against guessing, yielding conditions that should have minimized false recall. In addition, unlike other false memory paradigms, there was no attempt to explicitly insert misleading or false information (e.g., Loftus & Palmer, 1974). Nonetheless, the levels of false memories (as indexed by both objective and subjective measures) were among the strongest ever reported in the literature. Subjects reported that the false memories for critical items were quite compelling when tested either with a confidence judgment procedure or with Tulving's (1985) "remember"/ "know" procedure.

Perhaps because the results were so striking, many researchers attempted relatively direct replications of Roediger and McDermott's (1995) experiments (e.g, Payne, Elie, Blackwell, & Neuschatz, 1996; Schacter, Verfaellie, & Pradere, 1996, among many others) seeking to confirm (or not) their findings. The basic results are easily replicable, and in some sense, the findings from the DRM (Deese–Roediger–McDermott) procedure seem to be considered commonplace now, seven years after the original report. The prevailing wisdom seems to be that "Of course people recall, recognize and recollect the critical word that is related to all the words on the list. How could it be otherwise?" The interpretation of the results reported by Roediger and McDermott (1995) have been questioned on occasion (e.g., Miller & Wolford, 1999; but see Roediger & McDermott, 1999; Wickens & Hirshman, 2000; and Wixted & Stretch, 2000 for counterarguments), but everyone reports strong effects of false recall and false recognition under the conditions that Roediger and McDermott (1995) used.

Researchers now use the DRM paradigm to ask all sorts of interesting questions about the arousal of these types of false memories, generally using the 24 lists Roediger and McDermott (1995) used or the expanded set of 36 lists published in a norming study by Stadler, Roediger, and McDermott (1999). In fact, because these lists produce such robust false remembering, they have fuelled a cottage industry aimed at reducing the effect. Several manipulations have been identified that can reduce false remembering relative to true remembering (but rarely is the effect eliminated). Such manipulations include fully debriefing subjects about the illusion, and warning them to avoid false memories (e.g., Anastasi, Rhodes, & Burns, 2000; Gallo, Roberts, & Seamon, 1997; Gallo, Roediger, & McDermott, 2001b; McDermott & Roediger, 1998; Neuschatz, Payne, Lampinen, & Toglia, 2001; see also Libby & Neisser, 2001); presenting the list items in a distinctive format such as pictures, anagrams, or visual relative to auditory presentation (e.g., Gallo, McDermott, Percer, & Roediger, 2001a; Hicks & Marsh, 1999; Israel & Schacter, 1997; Smith & Hunt, 1998); or giving subjects repeated exposures to the study materials (e.g., Benjamin, 2001; McDermott, 1996; Seamon et al., 2002). Many of these investigations have been focused on the ability of the rememberer to control their memory accuracy via heuristics or monitoring strategies. For instance, the reductions in false remembering due to presenting lists in distinctive formats have been attributed to enhanced source monitoring (Hicks & Marsh, 1999) or other retrieval-based heuristics that subjects can use (Israel & Schacter, 1997; Schacter, Israel, & Racine, 1999). Similarly, warnings are most effective when given before study (relative to after study but before test), when subjects can strategically determine the critical item to some lists and consciously avoid misattributing these thoughts to actual presentation (Gallo et al., 2001b). By emphasizing cognitive control as a means to reduce memory errors, these findings complement those from several other paradigms that demonstrate that recollection-based control processes can keep familiarity-based errors in check (e.g., Brainerd, Reyna, & Mojardin, 1999; Jacoby, 1991; McElree, Dolan, & Jacoby, 1999; Rotello, Macmillan, & Van Tassel, 2000; Yonelinas, 1997).

Such approaches are theoretically important, and are useful for developing practical means by which false remembering can by minimized. However, by emphasizing individual or cognitive control over false memories they run the risk of overlooking the potentially powerful influence of stimulus control (see Watkins, 1991, for a discussion). In the context of the DRM procedure, this approach belies appreciation of a factor that has been relatively neglected, viz., that of variability of the potency of lists in producing false recollection. Roediger and McDermott (1995) did not use all of the 36 lists that Deese (1959) used in his experiments in their own studies, for the good reason that many of Deese's lists did not produce very high levels of false recall. (Deese never examined false recognition.) Only six lists were used in Roediger and McDermott's first experiment and 19 of Deese's lists were among their 24 used in Experiment 2. Several of Deese's other lists rarely produced false recall, with a critical intrusion rate at or below 10%.

The present experiments were conceived to address a puzzle inherent in Deese's (1959) original work: How can lists that are produced in exactly the same manner (the first 15 associates of a critical word that is not presented) produce either very high or very low levels of false recall? As a concrete example, consider the study lists corresponding to the critical items *bitter* and *sweet*. The 15 associates (Russell & Jenkins, 1954) to *sweet* that were used by Stadler et al. (1999) were *sour*, *candy*, *sugar*, *bitter*, *good*, *taste*, *tooth*, *nice*, *honey*, *soda*, *chocolate*, *heart*, *cake*, *tart*, and *pie*; whereas the 15 associates to *bitter* that were used in Experiment 1 of the present article were *sweet*, *sour*, *taste*, *chocolate*, *rice*, *cold*, *lemon*, *angry*, *hard*, *mad*, *acid*, *almonds*, *herbs*, *grape*, and *fruit*. Although both lists seem to converge on the meaning of their critical item and both lists seem to have face validity in producing the effect, the former is one of the best lists in terms of eliciting false recall (mean = .54), whereas we found the latter to be one of the worst (.01). Such comparisons should be made with caution because these data were obtained at different universities, but the differences are so great that we felt a more thorough investigation of this variability was warranted.

We set out to replicate the other, relatively unknown, half of Deese's work—that some lists of associates fail to elicit false recall—and to see if the pattern would extend to recognition measures. Relative to recall, recognition may depend more on familiarity than recollection (according to two-process recognition models such as those of Mandler (1980) and Jacoby (1991)) and therefore lists that produce negligible levels of false recall might still produce reasonably high levels of false recognition via a similarity-based familiarity mechanism. Although examination of types of materials has been relatively neglected, we believe that the theoretical stakes are high because some theories can account for such variability more easily than other theories.

One theory is that the theme or gist of the list is encoded, and the critical item is subsequently remembered because it is consistent with this gist representation (e.g., Payne et al., 1996, cf. Brainerd, Wright, Reyna, & Mojardin, 2001; Reyna & Brainerd, 1995). Here the critical item need not be activated prior to retrieval, but instead it is falsely remembered because it is consistent (i.e., has semantic overlap or is similar in meaning) with those items that were studied, and thus is highly familiar.[1] Another theory is that semantic features encoded from list items (the "exemplars") overlap with those of the critical item (the "prototype"), and this overlap leads to familiarity and hence false remembering (e.g., a semantic feature-matching theory; Arndt & Hirshman, 1998; cf. Anisfeld & Knapp, 1968). This theory differs from gist-based accounts at the representational level, in that the feature theory postulates that only individual exemplars need be stored (e.g., Hintzman, 1986, 1988) while fuzzy trace theory postulates a separate gist representation. At the functional level, however, the two accounts are similar in that both postulate that overlapping semantic features between the studied items and the critical item can yield false remembering. These theories make sense in the context of most DRM studies, where only those lists that elicit high levels of false remembering have been used, but a potential difficulty arises when one considers the variability among these lists to elicit the illusion. Certain lists produce negligible amounts of false recall, even though they were constructed (and appear) to converge on the meaning of their critical item. Hence, after studying these lists, there should be a sufficient number of overlapping semantic features in memory to drive false remembering.

As an alternative explanation, we have developed an activation/monitoring framework to account for false memory effects in the DRM paradigm and more generally (see Roediger, Balota, & Watson, 2001b for discussion of activation processes and Roediger, Watson, McDermott, & Gallo, 2001c for discussion of activation/monitoring). In brief, processing the list items (at study or test) activates the critical nonpresented

[1] The notion here is that semantic overlap exerts an automatic or nonconscious influence at retrieval, resulting in the subjective experience of familiarity. If the subject instead deliberately determined semantic overlap for a test item, then the subjective experience might be more like a decision process than remembering (see Whittlesea & Leboe, 2000 for a discussion). Elsewhere we have demonstrated that such explicit decision processes, conceptualized as item-specific criterion shifts, do not cause DRM false remembering (Gallo et al., 2001b).

associate, and false remembering reflects a failure to correctly monitor the source of this activation. This activation could be automatic spreading activation within the semantic network (e.g., Seamon, Luo, & Gallo, 1998; cf. Collins & Loftus, 1975) and/or conscious thought of the item due to more explicit associations (e.g., McDermott, 1997; cf. Underwood, 1965). In either case, it can lead to false remembering when the subject mistakenly attributes this activation to the item's occurrence during study (e.g., a breakdown in the reality monitoring process; Johnson & Raye, 1981; Johnson, Hashtroudi, & Lindsay, 1993). Note that the difference between this activation explanation and the semantic feature overlap explanation is nontrivial. If the former were true then the DRM effect would be a laboratory analog of falsely remembering internally generated events as having been externally perceived. If the latter were true then the effect would be analogous to falsely remembering events that were not previously encountered (internally or externally), but that were consistent with one's understanding (i.e., the meaning) of actually perceived events.

As discussed in the context of Experiment 1, one explanation of the list-level variability to elicit false recall is associative strength. Although all DRM lists consist of the 15 highest associates to a critical (nonstudied) item, as found in free association norms, the list items may vary in their tendency to elicit the critical item. This factor of "backward associative strength" should be critical according to activation/monitoring theory, and indeed the *sweet* list has high backward associative strength whereas the *bitter* list is relatively low on this dimension. It is less clear how such a relationship would be handled by gist or semantic feature-based theories, although the three are not mutually exclusive. We will discuss the relation of these theories more thoroughly in light of the results of the present experiments.

Experiment 1 is an exploratory study in which we attempted to verify Deese's (1959) failure to find false recall after presentation of certain associative lists and see if this pattern, if obtained, would extend to recognition procedures. Experiment 2 was designed to confirm and extend the results of Experiment 1 and to examine a variable—presentation rate—that has produced interesting and somewhat different results in prior DRM experiments. In brief, slowing the presentation rate of words beyond one second tends to decrease false recall (McDermott & Watson, 2001; Toglia & Neuschatz, 1996), but has little or no effect on false recognition (Arndt & Hirshman, 1998). We found parallel effects of presentation rate on false recall and recognition, but recognition always followed recall in Experiment 2. Experiment 3 was designed to overcome this limitation by omitting recall tests, and allowed us to directly compare two types of metamemory judgments (confidence and "remember"/"know"/"guess") for false memories as a function of type of list and presentation rate. The package of experimental results is discussed in terms of how associative factors and monitoring processes determine false memories and provides important constraints for current theories of false memories of this sort.

Experiment 1

Deese (1959) proposed that recall intrusions varied across lists because they differed in associative strength, or the degree to which all of the list words were related to each other and to the critical nonpresented word. In support of this claim, he found that the mean backward associative strength of a list (MBAS) was predictive of the likelihood of that list to elicit false recall of the critical item. BAS is the probability that a list item will evoke the critical item as a response in a free association task. The MBAS of a list is simply the average of all the backward associations between each list word and the critical item. Deese found that the probability of a critical item's intrusion was highly correlated with the MBAS of the study list ($r = +.87$). Thus, the more likely list items were to elicit the critical item on a free association test, the greater the probability of false recall from that particular list.[2]

[2] Robinson and Roediger (1997) found that the total backward associative strength to the critical lure (TBAS) was the critical predictor of false recall, rather then MBAS. TBAS is the sum of each list item's probability of eliciting the critical item on a free association task (whereas MBAS is TBAS divided by the number of items in the list). Because Deese (1959) held list length constant (as we do in the present experiments), the MBAS of a list was simply a linear transformation of the TBAS. In Robinson and Roediger, increasing the number of associates in a list (thereby increasing both MBAS and TBAS) enhanced false recall, but adding unrelated items to a list (thereby decreasing MBAS but not TBAS) did not affect false recall. Of course, both studies indicate that the degree of associative relationship from list items to the critical item determines the probability of false recall, and this is the main point for present purposes.

These results were extended by McEvoy, Nelson, and Komatsu (1999) who found that lists specifically created to have high BAS evoked greater false recall and false recognition than those created to have low BAS, although their lists were constructed in a different fashion than those of Deese (1959) and recognition was confounded with prior recall.

In this experiment we gathered normative recall and recognition data for lists which we thought would elicit low levels of false recall and false recognition (hereafter the "weak" lists). The procedures employed were closely modeled after the norming study of Stadler et al. (1999). All lists in the present experiment were created using procedures similar to those of Deese (1959) and Roediger and McDermott (1995), and consisted of the 15 words most frequently given to a nonpresented critical item on a free association test. The key feature was that these lists had low MBAS, which we expected would result in low levels of false remembering. As in Stadler et al. (1999), each list was followed by an immediate free recall test and a final recognition test was given at the end of the experiment.

Method

Subjects

The subjects were 80 Washington University undergraduates who participated for class credit after giving informed consent.

Materials

There were 28 study lists of 15 words each, all of which are in Appendix A. All lists consisted of the 15 highest associates to a critical nonpresented item, as found in the Russell and Jenkins (1954) word association norms. As in Roediger and McDermott (1995), words within a list were arranged in order of decreasing relatedness to the critical item. Relatedness was indexed by forward associative strength (FAS), which is the probability that the critical item elicited the list item on the free association test. Care was taken so that words were not repeated across lists, very similar words were not used within a list (e.g., *scene* and *scenery* were both listed as high associates to *beautiful*, but only one was chosen for the study list), and critical items were not used as study items in any of the lists. In all of these instances, the unwanted item was removed from the list and the highest unused associate from the Russell and Jenkins (1954) word norms was added to the end of the list. Other than these constraints, all lists used in this experiment were created by following Roediger and McDermott's (1995) procedures.

Twenty of the 28 lists were used because we felt they would elicit relatively low levels of false recall. Of these 20 lists, one (the *king* list) was chosen from the Stadler et al. (1999) set because it had produced the least false recall and false recognition of all the lists they tested. Similarly, seven lists (the *beautiful*, *butterfly*, *carpet*, *command*, *mutton*, *whistle*, and *wish* lists) were chosen from the Deese (1959) set, because Deese found that they elicited low levels of false recall relative to the other lists he used. Because Deese's lists consisted of 12 items, we added the next three highest associates from the Russell and Jenkins (1954) word norms to each of these lists to make them 15 items long. The final 12 lists used to elicit low levels of false memory were created from the Russell and Jenkins norms in a manner consistent with Roediger and McDermott (1995), and with the constraints outlined above. These particular lists (the *bitter*, *cabbage*, *citizen*, *cottage*, *health*, *justice*, *lamp*, *long*, *stove*, *swift*, *trouble*, and *whiskey* lists) were chosen because they had relatively low MBAS (equal to or less than .04), and the critical items did not seem to us to be orthographically distinct or uncommon relative to the critical items of other lists. According to Deese's (1959) linear regression [False recall $= 3.7 + 1.63 \times$ MBAS], lists with MBAS of .04 should elicit levels of false recall of approximately 10%. Thus, these 12 lists were expected to elicit relatively low levels of false recall. The average MBAS across all of the 20 lists chosen to elicit low levels of false recall was .028 ($SD = .048$), whereas it was .211 ($SD = .088$) for the 20 lists in Stadler et al. (1998) that produced the greatest false recall.

The BAS and FAS between each list item and the critical item were obtained via the Nelson, McEvoy, and Schreiber (1999) word association norms. However, of all 420 study items (28 lists × 15 items/list), 59 were not found in the Nelson et al. norms. We therefore gathered our own free association data for these items using similar procedures to those used in Nelson et al. (1999). These norms were part of a larger norming project consisting of at least 86 observations per item (see Roediger et al. 2001c, for details). These normative data supplemented those of Nelson et al. (1999), and the last two columns of Table 1 present the resulting MBAS and MFAS measures calculated for each list. Note that because list length was held constant, MBAS and MFAS are simply total associative strengths (TBAS and TFAS, respectively) divided by 15.

Table 1
Percentage of subjects who recalled list items (serial positions 1–15) and critical items (CI) for each list in Experiment 1

List	CI	Serial position															M	MFAS	MBAS
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
*Rough	54	86	67	47	18	81	34	15	44	52	35	42	77	68	77	86	55	.033	.116
*Needle	45	70	63	69	56	53	34	54	65	60	25	55	77	74	81	86	61	.062	.192
*Smoke	31	96	71	46	51	59	44	89	38	31	54	51	50	69	71	88	61	.025	.136
Justice	30	63	69	69	73	36	48	30	54	31	54	60	63	54	86	98	59	.044	.026
Wish	29	81	79	86	53	30	26	51	40	40	29	53	65	65	68	100	58	.053	.012
*Foot	25	79	70	71	75	56	91	66	21	64	49	61	48	78	81	90	67	.048	.173
*Trash	25	93	63	68	51	43	30	29	33	59	31	38	65	90	88	90	58	.058	.151
*City	24	81	90	63	49	45	69	39	95	39	81	49	98	66	65	93	68	.041	.178
*Rubber	21	80	68	43	41	60	33	39	23	51	43	65	50	54	76	95	55	.017	.033
Stove	18	89	86	53	53	50	45	56	34	60	31	49	55	68	83	91	60	.058	.035
Carpet	15	76	70	61	76	41	33	28	16	40	74	59	76	80	98	93	61	.039	.037
Lamp	14	80	70	63	54	34	49	36	59	26	59	65	58	79	78	100	61	.063	.006
*Pen	14	88	70	69	56	56	40	69	61	51	38	56	55	86	78	94	64	.054	.169
Swift	14	84	78	81	94	49	36	64	38	39	78	43	25	70	71	94	63	.058	.006
Health	11	95	46	70	61	64	81	34	24	81	44	28	49	68	89	98	62	.028	.020
Citizen	10	99	78	59	39	46	74	43	61	75	56	60	79	85	90	78	68	.046	.003
Whistle	9	76	84	54	43	33	25	23	55	38	64	90	70	75	78	93	60	.038	.005
Command	8	75	79	63	48	19	51	53	69	53	34	34	66	50	90	81	58	.035	.009
Cottage	8	84	78	63	64	68	64	46	38	55	54	50	69	71	85	96	66	.053	.003
Trouble	8	75	66	40	23	39	35	74	44	70	44	55	73	54	81	98	58	.026	.010
*King	6	95	80	65	88	78	61	50	34	46	59	56	78	70	78	86	68	.059	.224
Cabbage	5	88	96	66	28	63	76	66	51	65	38	41	48	56	95	96	65	.051	.011
Beautiful	3	95	93	93	81	76	58	36	40	71	88	59	65	41	80	94	71	.049	.038
Long	3	86	63	68	74	20	23	31	56	41	58	80	73	71	70	91	60	.045	.039
Whiskey	3	99	97	90	56	76	47	60	47	43	59	61	71	91	83	96	72	.042	.022
Bitter	1	99	96	68	79	65	31	64	60	48	45	63	71	44	83	99	68	.059	.011
Butterfly	1	85	89	69	75	64	80	41	48	30	76	85	49	85	83	84	70	.033	.045
Mutton	1	93	85	63	79	56	68	48	41	63	66	45	55	45	79	96	65	.014	.002
Mean	16	85	76	65	58	52	49	48	46	51	52	55	63	68	81	92	63	.044	.061

Note. The rate of noncritical intrusions across lists was 33%. Lists taken from Stadler et al. (1999) are denoted with an asterisk (*). M, mean recall of items 1–15; MFAS, mean forward associative strength; MBAS, mean backward associative strength.

The remaining eight study lists used in Experiment 1 were chosen from the Stadler et al. (1999) norms. We included eight of these lists so that we could compute the correlation between our recall and recognition data and those reported by Stadler et al. (the *city*, *foot*, *needle*, *pen*, *rough*, *rubber*, *smoke*, and *trash* lists), to ensure that our procedures would lead to comparable results. These eight lists were chosen because they had the fewest overlapping items with the other lists used in the present experiment. The average probability that these lists elicited false recall and false recognition in Stadler et al. was relatively high (.45 and .69, respectively). A total of nine words from these lists was replaced due to the constraints outlined above, and the average MBAS for the resulting set of lists was .144 ($SD = .051$). We chose an additional list (the *lion* list) from Stadler et al. to be used as a practice list. This list was always presented to subjects as the first list to be studied and recalled, and the resulting data were not analyzed. Across all 28 study lists, the average MBAS was .061 ($SD = .072$).

After all lists had been studied and recalled, a final recognition test was given. The recognition test consisted of 168 items. Half of these items were presented during the study phase and half were not. As in Roediger and McDermott (1995), test items were sampled from serial positions 1, 8, and 10 of each of the 28 study lists, in addition to the 28 critical lures. The remaining 56 items were randomly chosen words from the Nelson et al. (1999) norms which were unrelated to any of the list items. Test items were randomly arranged in two columns on test sheets, with the exception that words from any given list would be at least two items apart. The same test was given to all subjects.

Design

Because the recognition test was not given until all lists had been studied and recalled, the amount of time between the study of a list and the final recognition test was considerably greater for those lists studied early in the session compared to those studied later. To compensate for this we created four counterbalancing conditions with 20 subjects in each. With the exception of the practice list, the 28 study lists were randomly arranged and then assigned numbers. One subject group studied the lists in order (list 1–list 28), and this study order was reversed for a second group (list 28–list 1). A third group studied the lists in an inverted order (list 14–list 1, and then list 28–list 15), and this order was reversed for a fourth group. Therefore, across subjects, each list was presented an equal number of times in the beginning, middle, and end of the study/recall phase.

Procedure

Subjects were tested in groups of 9–13. They were instructed that they would hear 29 lists of 15 words, and after each list they were to write down in any order as many of the words from that list as they could remember. All 29 study lists were recorded on cassette tape in a male voice at a rate of 1.5 s per word (stimulus onset asynchrony), with the prompt “next list” occurring before each list. It was emphasized that subjects should write down only those words that they were certain they remembered, and not to guess. After each list was presented, the experimenter pressed the pause button on the tape player, signaling the subjects to recall the list.[3] Each subject was given a response booklet with the pages numbered to correspond to the list to be recalled on that page. Subjects were given two minutes to recall each list and were then instructed to turn the page before the presentation of the next list.

Following the recall of the last list, subjects were given instructions for the recognition test. They were told that the test consisted of some words they had studied (old) and some words they had not (new). Next to each word was a four-point scale for them to rate their confidence that the word had or had not been presented, ranging from 4 (sure the item was old) to 1 (sure the item was new). Subjects were told to go through the test at their own pace, and were instructed not to guess. Following the recognition test, subjects were given credit, debriefed, and thanked for their participation.

Results and discussion

Unless otherwise noted, all results in the present paper are significant at the .05 level, two tailed.

Recall

Recall results are presented in Table 1, with lists arranged in descending order of false recall of the critical items. From the table it can be seen that the mean proportion of false recall of critical items ranged widely, from .54 for the *rough* list to .01 for the *bitter*, *butterfly*, and *mutton* lists. As expected, the mean probability of false recall from the 20 lists which we thought would elicit low levels of false remembering was only .10 ($SD = .08$), and over half of these lists yielded false recall that was even lower than this level. The mean number of all noncritical intrusions for this set of 20 lists was .33 ($SD = .12$) per list. In contrast to false recall, true recall did not differ nearly as much across lists, ranging from .72 for the *whiskey* list to .55 for the *rough* and *rubber* lists (collapsing across serial positions 1–15). For the 20 lists used to elicit low levels of false recall, true recall was .64 ($SD = .05$). The Pearson product–moment correlation between true recall and false recall was −.54 ($n = 28$, $p < .01$) across lists, indicating that lists which were better remembered tended to elicit less false recall. Noncritical intrusions also correlated negatively with true recall ($r = -.42$, $n = 28$, $p < .05$).

In general, there was a high correlation between our norms and those of previous studies. The correlation between the recall of list items (averaging across items 1–15) in our experiment and in Stadler et al.’s (1999) experiment was +.90 across overlapping lists ($n = 9$, $p < .01$). For these lists, the average proportion of list items recalled was .62 in our experiment and .60 in Stadler et al. With reference to false recall, the correlation between these two experiments was

[3] For a group of 10 subjects the experimenter prematurely hit the pause button during the presentation of the *whiskey* list, so the recall and recognition results for this list are based on 70 subjects. Also, one subject failed to recall the *rough* list, so the results for this list are based on 79 subjects.

also high, $r = +.80$ ($n = 9$, $p < .05$), although average false recall for these nine lists in the present experiment (.27) was lower than that found by Stadler et al. (.41). This difference in absolute levels of false recall must be interpreted with caution, though, because we replaced some of the items from Stadler et al.'s lists with lower associates in order to eliminate redundancy across lists (as discussed previously). Considering our norms and Deese's (1959) experiment, false recall correlated highly across the seven overlapping lists, $r = +.90$ ($n = 10$, $p < .01$). The mean probability of false recall across these lists was .09 in our experiment and .07 in Deese's experiment. True recall for these lists was not reported by Deese (1959). Finally, we obtained split-half correlations on our recall data using the same procedure as Stadler et al. Subjects were arbitrarily divided into two groups, and for each group we calculated the average probability of true and false recall for each list. Much like Stadler et al., our data was very consistent between these two groups: the split-half correlation for true recall was +.83 ($n = 28$, $p < .01$) across lists, and for false recall the correlation was +.81 ($n = 28$, $p < .01$) across lists. Taken as a whole, the consistency between our norms and the results of other studies suggests that our finding of considerably low levels of false recall for some lists cannot simply be attributed to different subjects or procedures.

In the present experiment, the correlation between false recall and MBAS was significant, $r = +.49$, $n = 28$, $p < .05$, replicating Deese's (1959) finding. However, our correlation between MBAS and false recall was not of the same magnitude as Deese's ($r = +.87$), and may have been lower because we had less MBAS variability among lists due to our list-selection criteria. In order to increase the range in MBAS, we pooled the data from Stadler et al.'s (1999) 36 lists with the data from the 19 lists used in the present experiment that were not redundant with Stadler et al.'s 36 lists. The resulting set of 55 lists had average false recall of .30 ($SD = .20$) and average MBAS of .125 ($SD = .103$). On this expanded set, the correlation between false recall and MBAS increased to +.73, $n = 55$, $p < .01$. MBAS did not correlate significantly with true recall, and MFAS did not correlate significantly with true recall or false recall (all p's $> .05$).

Recognition

Recognition results are presented in Table 2, with lists arranged in descending order of overall false recognition of the critical item (i.e., a rating of 3, "probably old," or 4, "sure old"). Similar to the recall results, false recognition of critical items ranged from extremely high (87% of subjects falsely recognized *rough*) to extremely low (only 11% of subjects falsely recognized *mutton*). For the 20 lists chosen to elicit low levels of false recall, the mean probability of false recognition (.46, $SD = .19$) was considerably lower than that of true recognition (.76, $SD = .07$), collapsing across serial positions 1, 8, and 10. Still, for nine of these lists subjects falsely recognized the critical item at least half the time. In addition, for all 28 lists false recognition of the critical item was at least as high as the false alarm rate to unrelated distracters (.07), and in most cases it was considerably higher. Thus, even for most of the lists that elicited low levels of false recall, false recognition was generally high. Of course, the recognition test was considerably delayed relative to the recall tests because it occurred after subjects had studied and recalled all 28 lists. The correlation between true recognition and false recognition across all 28 lists was not significant ($r = +.03$, $p > .05$), consistent with Stadler et al. (1999).

Like the recall data, our recognition data were consistent across subsets of subjects, as the split-half reliability for true recognition ($r = +.83$, $n = 28$, $p < .01$) and false recognition ($r = +.89$, $n = 28$, $p < .01$) were both very high. For the nine lists which were obtained from Stadler et al. (1999), the correlation between our true recognition data and those of Stadler et al.'s experiment was high ($r = +.90$, $n = 9$, $p < .01$). For these lists, the average proportion of list items recognized (collapsing across serial positions 1, 8, and 10) was .81 in the present experiment and .71 in Stadler et al. Although the hit rate was higher in our experiment than in Stadler et al., the false alarm rate to unrelated items was also higher in our experiment (.07) than in Stadler et al. (.03), suggesting that our subjects may have been slightly more biased to make "old" judgments, in general. As was the case with true recognition, false recognition for these nine lists correlated highly with Stadler et al.'s results ($r = +.98$, $n = 9$, $p < .001$), and the average false recognition for these lists in the present experiment (.67) was very similar to that obtained in Stadler et al.'s experiment (.64). Much like the recall data, the consistency between the two studies suggests that the relatively lower levels of false recognition for many of the lists in the present experiment were not due to different subjects or procedures.

Table 2
Percentage of subjects who recognized list items (serial positions 1, 8, and 10) and critical items (CI) from each list in Experiment 1, and the corresponding percentage of high confidence recognition judgments

List	Overall recognition					High confidence recognition				
	CI	1	8	10	M	CI	1	8	10	M
*Rough	87	96	73	59	76	80	84	56	51	64
*Trash	84	98	84	68	83	68	94	69	53	72
Wish	80	89	68	39	65	70	79	60	28	55
Justice	76	85	88	50	74	49	80	68	40	63
*Rubber	73	94	44	78	72	52	86	35	63	61
*Smoke	70	99	78	83	86	49	98	69	75	80
Stove	70	91	74	51	72	53	83	61	45	63
*Foot	69	89	48	74	70	45	81	33	59	58
*Needle	69	97	90	63	83	56	84	80	51	72
*City	64	91	98	99	96	50	81	95	94	90
Lamp	63	94	78	75	82	51	79	66	61	69
Citizen	60	100	84	90	91	29	100	70	88	86
*Pen	59	96	76	61	78	41	89	70	56	72
Health	55	88	61	75	75	35	81	43	66	63
Command	55	75	74	50	66	33	65	66	40	57
Trouble	54	90	61	58	70	26	81	44	40	55
Whiskey	53	84	77	86	82	40	80	76	80	79
Carpet	49	89	48	88	75	36	76	26	81	61
Beautiful	44	96	66	94	85	36	95	59	91	82
Cabbage	44	75	63	58	65	34	64	58	35	52
Cottage	38	91	56	83	77	25	81	43	75	66
Swift	35	85	41	93	73	18	71	33	89	64
Long	34	90	76	68	78	24	81	69	59	70
Bitter	26	91	80	64	78	14	84	75	54	71
Butterfly	26	91	64	86	80	14	91	49	83	74
*King	25	99	80	68	82	18	95	63	58	72
Whistle	19	78	69	74	73	5	74	54	65	64
Mutton	11	96	66	62	75	4	91	53	54	66
Mean	53	91	70	71	77	38	83	59	62	68

Note. The mean false alarm rate to unrelated distracters was 7%, and the mean high confidence false alarm rate was 2%. Items taken from Stadler et al. (1999) are denoted with an asterisk (*). M, mean recognition of items 1, 8, and 10.

Unlike the positive correlation between false recall and MBAS, we failed to find a significant correlation between false recognition and MBAS ($r = +.30$, $n = 28$, $p > .05$). However, because restriction of range in MBAS may have attenuated this relationship, we again pooled our results with those of Stadler et al. (1999). Across the entire set of 55 lists, the correlation between false recognition and MBAS was significant, $r = +.43$, $p < .01$. Thus, as was the case with false recall, lists with higher associative strength yielded higher levels of false recognition, although this relationship could have been influenced by prior recall in the present design. MFAS did not correlate with false recognition, and neither MBAS nor MFAS correlated with true recognition (all p's > .05).

Confidence judgments

On the right of Table 2 are the percentage of subjects that recognized each test item with high confidence (i.e., a rating of 4, "sure the item was old"). Across lists, high confidence judgments were highly correlated to the "old"/"new" data, because a greater probability of recognizing an item as "old" would necessarily mean more opportunities for that item to have been recognized with high confidence. From the table it can be seen that, like the "old"/"new" recognition data, high confidence recognition of list items was relatively stable across lists (the mean across all lists was 68%, *SD* = 9%), while that to critical items varied considerably (from 80% for *rough* to 4% to *mutton*).

Of additional interest was whether there was between-list variability in the average confidence with which subjects falsely recognized a critical item. In other words, did some critical items tend to be falsely recognized with high confidence and others with low confidence, or was false recognition always equally compelling, even though it occurred more often for some lists? To address this issue, we again calculated the probability that a critical item was falsely recognized with high confidence, but expressed this probability as a proportion of only those instances where the item was falsely recognized (i.e., a rating of 3 or 4). This procedure is particularly informative because it expresses the probability of high confidence false recognition independent of the total proportion of false recognition, and as such it need not show differences across lists that vary in the total probability of false recognition. Nevertheless, the rate of falsely recognized critical items that were given a high confidence rating was 67% ($SD = 16\%$), averaged across lists, and ranged from 91% (for *rough*) to 27% (for *whistle*). The same rate for unrelated lures was 29% ($SD = 25\%$), which was at least as great as that for the list which elicited false recognition with the least confidence (i.e., *whistle*). Thus, even when only those instances in which a subject falsely recognized the critical item were considered, the persuasiveness of this memory illusion varied from list to list. Finally, the correlation across lists between the total probability of false recognition and the average confidence of a false recognition judgment was +.66 ($n = 28$, $p < .01$), suggesting that lists which elicited more false recognition also resulted in more compelling false recognition. The average confidence of a false recognition judgment also correlated with MBAS, $r = +.40$, $n = 28$, $p < .05$.

Relationship between recall and recognition

Like Stadler et al. (1999), we found significant positive correlations between true recall and true recognition, and also between false recall and false recognition. In the present experiment, the correlation between true recall and true recognition across lists was +.47 ($n = 28$, $p < .05$), and across subjects was +.72, ($n = 80$, $p < .001$). The correlation between false recall and false recognition was +.78 ($n = 28$, $p < .001$) across lists and +.61, ($n = 80$, $p < .001$) across subjects. However, as in Stadler et al., these correlations may have been influenced by the fact that every list was recalled before the final recognition test.

To summarize the main results, we found low levels of false recall for most of the lists with low MBAS (no greater than .04) that were created in the same fashion as those used by Roediger and McDermott (1995). We also found that lists with low MBAS resulted in relatively lower false recognition, although false recognition was still quite high relative to false alarms to unrelated distracters. Furthermore, the average confidence level accompanying false recognition from each list was positively correlated with the probability of false recognition from that list, and also with the MBAS of that list.

Experiment 2

The purpose of Experiment 2 was to confirm and extend the results of Experiment 1. To this end we examined the effects of presentation rate across lists which had been found to elicit either high or low levels of false memory in Experiment 1 and Stadler et al. (1999). Because presentation rate has been found to affect false remembering in this paradigm (as reviewed below), we wanted to determine if our set of lists that elicited very low levels of false recall would continue to do so at other presentation rates. Manipulating presentation rate would also provide theoretical insights into why some lists elicit high levels of false remembering while others do not. If the same mechanism underlies false remembering for these two types of lists, then presentation rate should have the same effects on false remembering from these lists.

Presentation rate is an interesting variable because there are two opposing predictions as to how presentation rate may affect DRM false memories. On one hand, one could argue that at slower rates subjects would have more time to encode information that could support false remembering. For instance, increased rehearsal of list items at slower rates might also increase the probability that the critical item would be thought of as an implicit associative response at study (Underwood, 1965). Similarly, at slower rates subjects would have more time to engage in semantic processing of the list items (e.g., encode semantic features), and to focus on the overall meaning of the list (e.g., the theme or gist). Evidence supporting these predictions is that deeper levels of processing at study, such as thinking of the meaning of the list items, have been found to increase false recall of critical items relative to more superficial processing (Thapar & McDermott, 2001; Toglia, Neuschatz, & Goodwin, 1999). In much the same way, slower presentation rates could lead to greater activation of the critical item, more thorough encoding of overlapping semantic features between the critical item and list items, and/or more thorough encoding of gist, each of which could increase false remembering.

On the other hand, by considering additional factors it could be argued that slower presentation rates would reduce the probability of false remembering. Even if slower rates resulted in more critical item activation or semantic feature/gist encoding, this would not necessarily lead to more false remembering. Rather, whether or not the critical item is falsely remembered would also depend critically on retrieval processes that operate on this information. Slow presentation rates would give subjects more time to engage in item-specific processing, resulting in more distinctive recollections of each individual item (e.g., Hunt & McDaniel, 1993). Distinctive processing of list items might facilitate editing processes whereby subjects are less likely to claim that they remember the critical item when memory for it is not nearly as compelling as memory for the list items (see Israel & Schacter, 1997; Schacter et al., 1999). In terms of the activation/monitoring theory, such editing processes are conceived as enhanced reality monitoring between memories for list items and critical items. Slowing presentation rate might enhance such a process (during study or test) and thereby reduce false remembering. In terms of gist-based theories, such editing processes could be conceived as increased reliance on specific traces of list items (e.g., verbatim traces in fuzzy trace theory) to counteract the effects of gist traces that can lead to false remembering (see Brainerd et al., 2001). In contrast, such editing processes are not necessarily inherent in feature-matching or exemplar-based models. Indeed, Arndt and Hirshman (1998) argued that MINERVA2 predicts increments in false recognition with greater degrees of list learning (i.e., at slower rates), not decrements.

Studies that have manipulated presentation rate have supported one prediction or the other, depending on the range of presentation rates used. Under conditions of very rapid presentation, increasing study time increased both true and false remembering. For instance, Roediger et al. (2001a, Experiment 1) used rapid serial visual presentation of DRM lists (ranging from 20 ms per item to 320 ms per item, with a 32 ms interstimulus interval) for immediate free recall. Recall of list items increased from .10 at 20 ms to .31 at 320 ms, and false recall followed the same pattern (.10 at 20 ms. to .33 at 320 ms). A similar trend was observed for true and false recognition (Experiment 3), although the effects were not significant. Seamon et al. (1998) also varied visual presentation duration of DRM lists to determine the effects on false recognition. They found that both true and false recognition (adjusted for baserate) increased as presentation duration was increased from 20 to 2000 ms, although the effect on false recognition was only significant in the second of two experiments. Considered together, these results suggest that under very rapid presentation conditions (i.e., half a second per item or less) neither the list items nor the critical item or theme is activated or encoded very well. Thus, compared to very rapid presentation conditions, slowing presentation duration increased false recall and recognition, which is consistent with an increase in activation/encoding of the information that causes false remembering.

Using slower presentation rates, Toglia and Neuschatz (1996) have demonstrated the opposite effect of presentation time. In their study, DRM lists were auditorily presented to subjects at interstimulus intervals of either 1 or 4 s for immediate free recall. Whereas accurate recall increased as presentation rate slowed (.65 compared to .71), false recall decreased (.72 compared to .49). These results are consistent with the second prediction outlined above, and suggest that subjects in the slower rate condition were better able to use editing processes to reduce false recall. That is, although the studies using rapid presentation suggest that activation/encoding of the critical item, overlapping semantic features, or gist accrues as presentation time is slowed, presentation time can apparently become sufficiently slow so that the enhanced memory for list items allows subjects to reduce false responding. Consistent with this claim, McDermott and Watson (2001) have demonstrated an initial increase and eventual decrease in false recall as a result of slowing visual presentation duration. Specifically, increasing presentation duration from 20, 250, 1000, 3000, and 5000 ms resulted in enhanced recall of list items (.17, .31, .41, .48, .49) but a nonlinear pattern for critical items (.14, .31, .22, .14, .14). It is less clear, though, if slowing presentation time also reduces false recognition. Arndt and Hirshman (1998) provided relevant data. Collapsing across relevant conditions in three of their experiments (weighting each equally), they found that slowing presentation duration resulted in increased false recognition (.59–.72) at rapid durations (300–800 ms, respectively), but further slowing did not have an impact (.72 at 3000 ms). Thus, increased study time increased false recognition to a point, but slowing the rate further did not affect false recognition.

In the present study we used a range of presentation rates within which we felt the critical item, semantic features, or gist could be strongly activated/encoded even at the fastest rate (.5, 1, and 3 s, inter-stimulus intervals; or approximately 1, 1.5, and 3.5 s SOA). As in Experiment 1, all study items were presented auditorily, and memory was tested using both immediate free recall and also a final recognition test after all lists had been presented. Based on previous research, we predicted that slowing presentation rate would result in decreased false recall for those lists that have been found to elicit high levels of false remembering, but we were unsure how rate would affect false recognition. It was equally unclear how lists that had elicited low levels of false remembering would react to presentation rate. Assuming that the same types of information cause false remembering from both sets of lists, albeit at quantitatively different levels, we predicted that presentation rate would affect false remembering from both sets of lists in a similar fashion. However, it was possible that these weakly related lists elicited very low levels of false recall in Experiment 1 because we did not provide enough study time for subjects to adequately grasp the theme or gist of the lists. If this were the case, then slower presentation rates would afford more time for subjects to note the theme of these lists, providing a greater chance for these lists to elicit high levels of false recall.

Method

Subjects

The subjects were 90 Washington University undergraduates who participated for class credit or for $8 after giving informed consent.

Materials

Twenty-four word lists were used in this experiment. Twelve of these lists were chosen from Stadler et al. (1999) because they had been found to evoke the highest levels of false recall in that study, and will be referred to as the strong lists (the *chair*, *city*, *cup*, *doctor*, *needle*, *rough*, *sleep*, *smell*, *smoke*, *soft*, *trash*, and *window* lists). The *sweet* and *anger* lists from Stadler et al. were not used because these lists overlapped too extensively with other lists used in this experiment. From Stadler et al., the average probability of false recall and false recognition of the critical items for these twelve lists were .54 ($SD = .07$) and .77 ($SD = .07$), respectively, and the average MBAS from these lists is .213 ($SD = 0.091$). The other 12 lists were those which we found to evoke the lowest levels of false recall in Experiment 1, and will be referred to as the weak lists (the *beautiful*, *bitter*, *butterfly*, *cabbage*, *citizen*, *command*, *cottage*, *king*, *long*, *trouble*, *whiskey*, and *whistle* lists). The *mutton* list was not included because this critical item is a very low frequency word and may be unfamiliar to many students. From Experiment 1, the average probability of false recall and false recognition of the critical items for these 12 lists were .05 ($SD = .03$) and .40 ($SD = .14$), respectively, and the average MBAS for these lists is .036 ($SD = .063$). Because we were not gathering normative data for each list in this experiment, we were not as concerned about the overlap of items between lists as we were in Experiment 1. Thus, any items that repeated between lists were not replaced, except for *smell* in the *cabbage* list, which was replaced with *rabbit* because of the occurrence of *smell* as a critical item for another list.

The two sets of critical items did not significantly differ in average printed word frequency (Kucera & Francis, 1967), the mean raw count per million was 108.75 for weak critical items and 81.83 for strong critical items, $t(22) < 1$. If anything, weak critical items were slightly higher in frequency, which typically leads to higher levels of false alarms in other paradigms (Glanzer & Adams, 1985).[4] The two sets of lists also did not differ in terms of interitem associative strength, or the average number of associative connections between list items (mean = 1.81 and 1.68 for weak and strong lists, respectively, $t(22) < 1$). This variable has been found to affect DRM false remembering in some studies (e.g., McEvoy et al., 1999), but not in others (e.g., Roediger et al. 2001c).

One master copy of the word lists was recorded in a female voice on an IBM-compatible computer

[4] As a follow-up analysis, we excluded the two most frequent critical items from each of the list sets (i.e., *city* for the strong lists and *long* for the weak lists), which were outliers. This resulted in average critical item frequency of 53.55 for the strong lists and 50 for the weak lists ($t(20) < 1$). Analysis of the data based on these attenuated list sets revealed the same pattern of results and conclusions in both Experiments 2 and 3.

using sound recording software. Different interstimulus intervals were then created by splicing various lengths of silence (.5, 1, or 3 s) between words on this master copy. Thus, the same recording was used across all presentation rates. The different versions of the master copy were then transferred to cassette tapes, allowing subjects to be tested in groups.

The recognition test consisted of 96 items that were randomly ordered on test sheets, with the exception that at least two words separated items from the same list. As in Experiment 1, half of the tested words had been presented during the study phase and half had not been presented. Test items included those words from serial positions 1, 8, and 10, as well as the critical items from each of the 24 lists. As discussed below, each subject studied 16 lists, yielding a total of 48 list items on the test. The 48 distracters were the 24 list items from the eight nonstudied lists, and the 24 critical items from all the lists. Because some items were repeated across different lists, care was taken to sample items that were unique to a given list. Due to this constraint, four items were chosen from serial positions other than 1, 8, or 10 (in these cases, the next item in order was sampled). All subjects received the same recognition test.

Design

A 3 × 2 mixed factorial design was used, with presentation rate (.5, 1, and 3 s) manipulated between subjects, and list type (strong vs. weak) manipulated within subjects. The study lists were divided into three sets of eight, each containing four strong lists and four weak lists which were randomly ordered within a set. At each of the presentation rates, there were three counterbalancing conditions to control for the possibility of list study-order effects. In each counterbalancing condition, subjects studied two sets of lists (a total of 16 lists), and the nonstudied set was used for distracters on the recognition test. Accordingly, one group studied Set X followed by Set Y, another studied Set Z followed by Set X, and a third studied Set Y followed by Set Z. There were 10 subjects in each counterbalancing condition at each presentation rate, resulting in 30 subjects tested at each presentation rate.

Procedure

All procedures of Experiment 1 were followed as closely as possible. Subjects were tested in groups of 1–12, with presentation rate manipulated between subjects. They were instructed that they would be presented with 16 lists of 15 words each for immediate free recall, with instructions that were the same as those of Experiment 1. Because subjects clearly understood the task demands in Experiment 1, we opted not to use a practice list in this experiment. After the final list had been recalled, subjects were given the recognition test with the same instructions that were used in Experiment 1. Following the test, they were debriefed, given credit or paid, and thanked for their participation.

Results and discussion

Recall

The proportion of list items (collapsing across serial position) and critical items recalled at each presentation rate is graphically presented in Fig. 1. Strong list results are in the top panel (A)

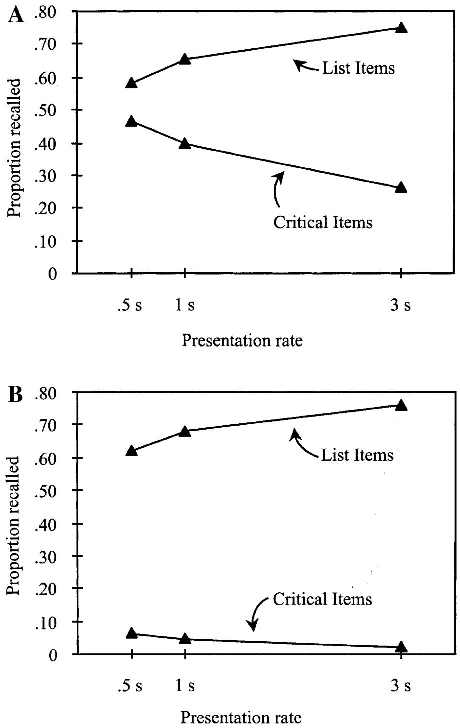

Fig. 1. Proportion of list items and critical items recalled from strong lists (A) and weak lists (B) in Experiment 2, as a function of presentation rate. For strong lists, the probability of true recall (across serial positions 1–15) was .59 (.5 s), .65 (1 s), and .75 (3 s), and the probability of false recall was .47 (.5 s), .40 (1 s), and .26 (3 s). For weak lists, the probability of true recall was .62 (.5 s), .68 (1 s), and .76 (3 s), and the probability of false recall was .07 (.5 s), .05 (1 s), and .02 (3 s).

and weak list results are in the bottom panel (B), and the actual proportions are listed in the caption. It is clear from the figure that slower presentation rates resulted in increased recall of the list items but decreased false recall of the critical items. This pattern was obtained for both the strong and the weak lists, consistent with the notion that the same underlying mechanism causes false remembering from these two types of lists. As expected, false recall for weak lists was much lower than that for strong lists at each presentation rate, while true recall did not differ greatly. Considering noncritical intrusions, a 2 (list type) × 3 (presentation rate) ANOVA indicated that there was no effect of list type, no effect of rate, and no interaction between the two (all F's < 1). Thus, unlike critical intrusions, noncritical intrusions did not differ from strong lists to weak lists and were not affected by presentation rate, although this may have been due to a floor effect. Collapsing across rates and list types, the mean number of all noncritical intrusions was .32, per list.

A 2 (item type) × 2 (list type) × 3 (presentation rate) ANOVA yielded significant main effects of item type, $F(1,87) = 674.29$, $MSE = .029$, as recall of list items was greater than false recall of critical items in every condition. There was also a main effect of list type, $F(1,87) = 191.32$, $MSE = .011$, and a significant interaction between list type and item type, $F(1,87) = 223.41$, $MSE = .013$. This interaction suggests that false recall was greater (by design) for strong lists (mean = .38) than for weak lists (.04), whereas true recall was relatively invariant (.66 and .69, respectively). The main effect of rate was not significant, but this variable did interact with both item type, $F(2,87) = 21.09$, $MSE = .029$, and list type, $F(2,87) = 3.42$, $MSE = .011$, and the three-way interaction between presentation rate, item type, and list type was also significant, $F(2,87) = 4.69$, $MSE = .013$. These interactions indicate that rate had opposite effects on true and false recall, and that the magnitude of the effects was different for strong and weak lists.

To evaluate the interactions between rate and list type, we computed separate ANOVAs on critical items and list items. A 2 (list type)× 3 (presentation rate) ANOVA on false recall of critical items indicated a main effect of list type, $F(1,87) = 220.66$, $MSE = .023$, demonstrating that false recall from strong lists (.38) was greater than that from weak lists (.04), collapsed across presentation rates. This replicates the key finding in Experiment 1 that there were sizable differences in ability of lists to elicit false recall. The main effect of presentation rate was also significant, $F(2,87) = 7.44$, $MSE = .035$, and presentation rate interacted with list type, $F(2,87) = 4.29$, $MSE = .023$. The interaction suggests that false recall from the strong lists was affected more by presentation rate than false recall from weak lists, although this was probably due to a floor effect for weak lists. Separate one-way ANOVAs demonstrated that the effects of rate on false recall were reliable for both the strong lists, $F(2,87) = 6.47$, $MSE = .005$, and weak lists, $F(2,87) = 3.59$, $MSE = .005$.

Similar analysis of true recall (collapsing across serial positions 1–15) revealed a main effect of list type, $F(1,87) = 17.24$, $MSE = .001$, demonstrating that recall was greater for weak lists (.69) than for strong lists (.66), collapsing across presentation rates. This finding is consistent with the finding in Experiment 1 that true recall correlated negatively with false recall across lists, although as in Experiment 1, true recall did not differ nearly as much as false recall across lists. The main effect of rate was also significant, $F(2,87) = 25.83$, $MSE = .014$, with recall increasing at slower rates, and this variable did not interact with list type, $F(2,87) = 1.07$, $MSE = .001$.

Serial position curves for strong lists at each presentation rate can be found in Fig. 2. These curves were smoothed by averaging adjacent points (e.g., the data point for serial position 4 represents the mean probability of recall for serial positions 3–5), with the exception of the endpoints (serial positions 1 and 15) which represent the probability of recall from those positions alone. Dashed lines represent the mean probability of false recall of the critical item. From the figure it can be seen that the recall of items in middle serial positions was affected by presentation rate to a greater degree than recall of items at the beginning (the primacy effect) or at the end (the recency effect) of the list, and this may have been due to a ceiling effect for these items. For these strong lists, false recall became indistinguishable from recall of list items presented in middle serial positions as presentation rate increased. Indeed, in the .5 s condition where false recall was greatest, recall of list items from the middle serial positions 4–11 (.44) was no greater than false recall of the critical items (.47), $t(29) = .49$, $SEM = .053$, $p = .63$, indicating robust levels of false remembering. In contrast, false recall from the weak lists was

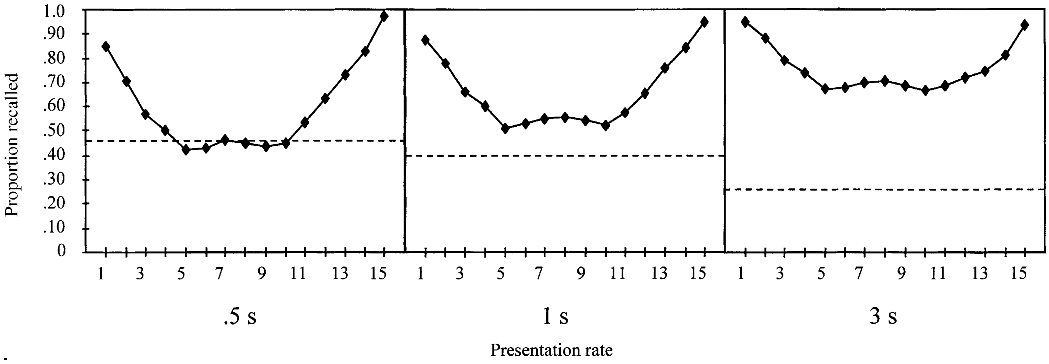

Fig. 2. False recall rates and smoothed serial position curves (see text) for strong lists in Experiment 2, as a function of presentation rate. Solid lines represent the probability of true recall across serial positions, and dashed lines represent the probability of false recall.

considerably low at every presentation rate, further demonstrating that these lists do not elicit high levels of false recall even under conditions where subjects should have had enough time to grasp the theme of these lists.

Recognition

Recognition data are presented in Table 3 as a function of list type (strong vs. weak) and presentation rate, and are divided into overall recognition and high confidence recognition, as in Experiment 1. In general, these data followed the same pattern as the recall data, although they must again be qualified with the fact that the recognition test always followed recall testing. First, false recognition of critical lures was higher for strong lists than weak lists in each condition, whereas true recognition did not differ nearly as much. Second, slowing presentation rate appeared to increase true recognition while decreasing false recognition for both strong and weak lists. Finally, false alarms to list and critical items from nonstudied lists (i.e., list

Table 3
Mean percentage of list items (collapsed across serial positions 1, 8, and 10) and critical items recognized in Experiment 2, and corresponding percent of high confidence judgments, as a function of list type and presentation rate

List type	Overall recognition Presentation rate			High confidence recognition Presentation rate		
	.5 s	1 s	3 s	.5 s	1 s	3 s
Strong lists						
List items						
Studied	81	84	87	68	73	81
Nonstudied	7	9	6	1	2	1
Difference	74	74	81	67	71	80
Critical items						
Studied	80	73	58	60	50	43
Nonstudied	8	9	13	3	3	1
Difference	72	64	45	57	48	42
Weak lists						
List items						
Studied	78	80	85	68	71	79
Nonstudied	8	9	9	1	3	4
Difference	70	71	76	67	68	75
Critical items						
Studied	39	41	24	19	25	14
Nonstudied	5	8	3	0	1	1
Difference	35	34	21	19	24	13

and critical item controls) were low and did not seem to be systematically affected by presentation rate. A 2 (item type) × 2 (list type) × 3 (presentation rate) ANOVA on these baserate false alarms indicated that there was a significant interaction between item type and list type, $F(1,87) = 5.24$, $MSE = .012$, whereas no other main effects or interactions were significant (all p's $> .05$). Follow-up analyses indicated that weak critical item baserates (mean = .05, collapsing across rates) were significantly lower than weak list item baserates (.09), $F(1,87) = 4.86$, $MSE = .007$, and strong critical item baserates (.10), $F(1,87) = 4.07$, $MSE = .089$. No other effects or interactions were significant (all p's $> .10$).

Such baserate variations potentially reflect idiosyncratic differences between item types. To adjust for these differences, these baserates were subtracted from the recognition rates of list items and critical items from studied lists (Schacter et al., 1996; Seamon et al., 1998). The residual false recognition effect (after adjustment) is considered that which is due to the presentation of the list of associates per se, in the absence of the influence of idiosyncratic item differences (or other influences) on false alarms. These adjusted data can also be found in Table 3. Except where noted, we report analyses only on adjusted scores. However, the same pattern of statistical results and general conclusions were obtained when unadjusted scores were analyzed.

A 2 (item type) × 2 (list type) × 3 (presentation rate) ANOVA on recognition of list items and critical items yielded a significant main effect of item type, $F(1,87) = 137.92$, $MSE = .056$, as true recognition was, on average, greater than false recognition. There was also a significant main effect of list type, $F(1,87) = 99.22$, $MSE = .027$, and an interaction between list type and item type, $F(1,87) = 71.01$, $MSE = .023$, again suggesting that false recognition differed considerably between the two types of lists, whereas true recognition did not differ as much. The main effect of presentation rate was significant, $F(2,87) = 3.79$, $MSE = .042$, and this variable interacted with item type, $F(2,87) = 10.61$, $MSE = .056$, suggesting that slower rates decreased false recognition while increasing true recognition. No other interactions were significant (all p's $> .10$).

We again computed separate ANOVAs on critical items and list items. Considering the critical items first, a 2 (list type) × 3 (presentation rate) ANOVA revealed a significant main effect of list type, $F(1,87) = 108.59$, $MSE = .039$, demonstrating that false recognition from strong lists (mean = .60) was greater than that from weak lists (.30), collapsing across rates. There was also a significant effect of rate, $F(2,87) = 10.12$, $MSE = .067$, and this variable did not interact with list type ($p < .10$). Thus, slowing rates decreased false recognition for strong and weak lists in a similar fashion. For the list items, there was a main effect of list type, $F(1,87) = 6.04$, $MSE = .011$. This indicates that true recognition from strong lists (.76) was greater than that from weak lists (.73), collapsing across rate. This is the opposite pattern than was found in true recall, although the effects in both instances were small. The main effect of presentation rate did not reach significance, $F(2,87) = 2.37$, $MSE = .031$, $p = .10$, and this variable did not interact with list type ($F < 1$). As opposed to the large effects of rate on false recognition, slowing rates did not affect true recognition that much (albeit in the expected direction).

Finally, recognition performance for the strong lists in the .5 s condition demonstrated a robust false memory effect: here the recognition rate of the list items (.74) was no different than that of the critical items (.72), $t(29) = .420$, $SEM = .051$, $p = .68$. Indeed, as can be seen from Table 3, in all the conditions of this experiment false recognition of the critical lures from studied lists was higher than recognition of the comparable items from nonstudied lists. Even in the condition in which false recognition of critical items was lowest (weak lists at the 3 s rate), unadjusted false recognition of critical items (mean = .24) was still considerably higher than false alarms to critical items of nonstudied lists (.03), $t(29) = 6.02$, $SEM = .035$.

Confidence judgments

On the right of Table 3 are the high confidence recognition data, which were also adjusted by subtracting baserate false alarms. As in Experiment 1, these data necessarily overlapped highly with the "old"/"new" recognition data. A 2 (list type) × 3 (rate) ANOVA on critical items revealed that those from strong lists (mean = .49, collapsing across rates) were falsely recognized more often than those from weak lists (.19), $F(1,87) = 161.57$, $MSE = .026$. The main effect of rate was marginal, $F(2,87) = 2.71$, $MSE = .069$, $p = .07$, and the interaction between list type and rate was significant, $F(2,87) = 3.04$, $MSE = .026$. These effects indicate that high confidence false recognition generally decreased with slower rates, although the effect was more pronounced for strong lists. Similar analysis of list items demonstrated that high confidence recognition was slightly greater for strong than weak lists (mean $=.73$ and .70, respectively), $F(1,87)=3.80$, $MSE=.010$, $p=.06$. These ratings increased with slower presentation rates, $F(2,87)=5.85$, $MSE=.031$, and there was no interaction between list type and rate, $F<1$. Finally, as with the "old"/"new" data, high confidence recognition of list items and critical items from nonstudied lists was considerably lower than that of list items and critical items from studied lists (all p's $<.01$).

As in Experiment 1, we calculated the proportion of recognized items that were given high confidence judgments. This was done only on those items associated with studied lists (i.e., unadjusted data), because items from nonstudied lists were not recognized often enough to yield a meaningful analysis. Also, because some subjects did not falsely recognize any critical lures from the weak lists, particularly in the 3 s condition (five subjects), unequal numbers of observations precluded statistical analysis on these ratings at the subject level. Thus, the means and analyses reported here were calculated across items instead of across subjects. At each presentation rate, the proportion of falsely recognized critical items given high confidence ratings from strong lists was .75 (.5 s), .67 (1 s), and .75 (3 s), and for weak lists was .56 (.5 s), .58 (1 s), and .61 (3 s). A 2 (critical item type) × 3 (presentation rate) ANOVA revealed that there was a main effect of list type, $F(1,33)=9.30$, $MSE=.041$, with more high confidence ratings given to false recognition judgments from strong lists (mean $=.72$) than from weak lists (.58). This finding is consistent with results of Experiment 1, in which it was found that the probability of false recognition from a list correlated positively with the mean confidence ratings given to false recognition from that list. Neither the main effect of presentation rate nor the interaction between rate and critical item type was significant (both p's $>.10$). Thus, although both list type and rate modulated the magnitude of the false recognition effect, only the former affected confidence ratings for false recognition. This dissociation bolsters the claim that different processes underlie the effects of list type and rate on false memory, a point to which we return in the General discussion.

In contrast to the confidence ratings to falsely recognized critical items, the confidence ratings given to correctly recognized list items were affected by presentation rate. At each presentation rate, the proportion of high confidence ratings given to correctly recognized list items from strong lists was .82 (.5 s), .86 (1 s), and .93 (3 s), and for weak lists was .86 (.5 s), .89 (1 s), and .92 (3 s). A 2 (list type) × 3 (presentation rate) ANOVA across items revealed only a main effect of presentation rate, $F(2,105)=9.02$, $MSE=.015$, with no main effect of list type and no interaction between the two (both p's $>.10$). The effect of presentation rate suggests that subjects recognized list items with more confidence as presentation rate slowed, as would be expected if subjects had engaged in more elaborative, item-specific processing at slower rates.

In sum, the findings of Experiment 2 demonstrate that both false recall and false recognition decreased as presentation rate was slowed from .5 s to 3 s (inter-stimulus interval). Furthermore, it was shown that presentation rate had the same effect on false remembering from both strong lists and weak lists, suggesting that the same mechanisms cause false remembering from these two sets of lists. We also replicated Experiment 1 by showing that strong lists resulted in greater false recall and false recognition than weak lists, while true remembering did not differ as much between these two types of lists. Of course, the recognition results were confounded with prior recall, an issue we address in the next experiment. Finally, subjects were equally confident in recognizing list items from strong and weak lists, whereas they were more confident in falsely recognizing critical items from strong lists than weak lists. Confidence judgments for recognized list items increased at slower rates, but confidence judgments for falsely recognized critical items were not affected by rate.

Experiment 3

The purpose of Experiment 3 was to explore list level differences in false recognition when the recognition test is not confounded with prior recall. Roediger and McDermott (1995) found a testing effect on true and false memories, so that prior recall of lists enhanced subsequent recognition of list items and critical items, although the effect has not been universally obtained in other studies (see Roediger, McDermott, & Robinson (1998), for a discussion). In those instances where the effect is found, the facilitation of both true and false recognition may be due to similar mechanisms, such as retrieval practice or the additional encoding opportunity provided by prior recall (cf. Tulving, 1967). However, it is unknown if these mechanisms will affect strong and weak lists equivalently, so that the causes of list-level differences in false recognition of Experiments 1 and 2 are unclear.

We addressed this issue in the present experiment by using the same materials and design as Experiment 2, with only two exceptions. First, during the 2 min between each study list's presentation subjects completed math problems instead of recalling the list. We could therefore be confident that any list-level differences in the present experiment are not due to prior recall. This design also allowed us to determine if prior recall was responsible for the effects of rate on false recognition in Experiment 2. Slower rates decreased false recognition in Experiment 2, but Arndt and Hirshman (1998) did not find consistent reductions in false recognition at slower rates (in the absence of prior recall), suggesting that our effects may have been driven by prior recall. Second, all subjects in Experiment 3 received the same recognition test as in Experiment 2, but half the subjects made confidence judgments to these items (as in Experiment 2), and the other half made "remember"/"know" judgments to these items. This was done to replicate the list-level variability in false recognition confidence judgments found in Experiment 1 and 2, and also to determine if these differences would extend to "remember"/"know" judgments. To anticipate, both subjective judgments yielded very similar patterns of results.

A final issue of interest is the level of false alarms to list items and critical items from nonstudied lists (i.e., nonstudied control items). In Experiment 2 these baserates were consistently low (at or below 13%), but they were significantly higher for critical item controls from strong lists than those from weak lists. This result was unexpected, because the corresponding lists were never studied and hence all of these items should have been relatively unfamiliar to the subjects. However, Roediger and McDermott (1995) found that critical item controls had higher baserates than corresponding list item controls, and similar effects were subsequently reported (e.g., Gallo et al., 2001a,b; Schacter et al., 1996; Seamon et al., 1998). The fact that we found baserate differences between critical item controls from strong and weak lists suggests either that these items have idiosyncratic differences, or that other factors influenced these baserates. We were therefore interested to see if such differences would be replicated in the present experiment, when no recall test occurred prior to the recognition test.

Method

Subjects

The subjects were 90 Washington University undergraduates who participated for class credit after giving informed consent.

Materials and design

As in Experiment 2, 30 subjects were randomly assigned to each presentation rate condition (.5, 1, or 3 s per word). The same word lists and counterbalancing procedures were employed as in Experiment 2. There were only two differences between this experiment and the last. The first is that subjects in the present experiment were each given a 16-page booklet of multiplication problems, so they could do math after each list instead of recall. The second is that half the subjects in each rate condition made confidence judgments to test items (using the same test as in Experiment 2), whereas the other half made "remember"/"know"/"new" judgments (explained below). Importantly, the two recognition tests had the same items arranged in the same random order, and the only difference was that decision choices to the right of each item was either a 1–4 confidence scale or an "R"/"K"/"New" judgment, respectively.

Procedure

Subjects were told that this was an experiment of memory and mathematics abilities. As in the previous experiments, they were told that they would hear 16 lists and that they should pay attention to each item in the list because their memory would be tested after all lists had been studied. After each list, they were told to turn their booklet to a new page of multiplication problems and to accurately complete as many problems as possible. They were asked to use this time only to focus on the math problems, and to avoid writing any of the list items on the math pages. No subject completed an entire page of math problems in the 2 min.

Following the last page of math problems, subjects received recognition test instructions. For those subjects making confidence judgments, these instructions were identical to those of the previous experiment. For those making "remember," "know," and "new" judgments the instructions were modified slightly (see Gardiner, 1988; Rajaram, 1993). Specifically, they were to make a "remember" judgment if the test word brought to mind a specific detail about the word's occurrence in the study list. "Know" judgments were to be made when the subject felt that they knew the word had been presented, but could not recollect any specific details about its actual occurrence. If the subject did not think the word was "old," i.e., they neither "remembered" nor "knew" that it had been presented, they were to circle "new."

Results and discussion

Recognition data for each item type are presented in Table 4 as a function of list type (strong vs. weak), test format (confidence vs. "remember"/"know" judgments) and presentation rate. The first point to notice is that, in all conditions, strong critical items were falsely recognized more often than weak critical items, replicating the results of Experiment 2. Strong list items were also recognized more often than weak list items, although as in Experiment 2, this difference was not nearly as large as that for critical items. With respect to presentation rate, recognition of list items generally increased at slower presentation rates, whereas false recognition of critical items did not appear to be systematically affected by rate. Finally, as in the previous experiment, baserate false alarms to strong critical items (i.e., those from nonstudied lists) tended to be somewhat higher than were those to weak critical items. A 2 (item type) × 2 (list type) × 2 (test format) ×3 (presentation rate) ANOVA on these baserate data indicated a main effect of list type, $F(1,84) = 18.42$, $MSE = .035$, and an interaction between list type and item type, $F(1,84) = 10.57$, $MSE = .03$. No other main effects or interactions were significant. Follow-up analyses revealed that baserates to strong critical items (.33) were greater than those to strong list items (.26), $t(89) = 2.48$, $SEM = .027$, and baserates to weak critical items (.19) were lower than were those to weak list items (.24), $t(89) = 2.32$, $SEM = .022$, and strong critical items, $t(89) = 4.53$, $SEM = .032$. There were no differences in baserate responding to list items from strong lists (.26) and weak lists (.24), $t(89) = 1.20$, $SEM = .021$.

Because of these baserate differences we report analyses on the adjusted data, although as in the previous experiment the same conclusions were

Table 4
Mean percentage of list items (collapsed across serial positions 1, 8, and 10) and critical items recognized in Experiment 3, as a function of list type, presentation rate, and test format

List type	Presentation rate								
	Confidence format			Remember/Know format			Mean		
	.5 s	1 s	3 s	.5 s	1 s	3 s	.5 s	1 s	3 s
Strong lists									
List items									
Studied	70	70	80	74	71	84	72	70	82
Nonstudied	30	18	22	32	29	27	31	23	24
Difference	40	52	58	42	42	57	41	47	58
Critical items									
Studied	82	86	76	82	83	85	82	85	80
Nonstudied	37	33	27	28	46	27	33	40	27
Difference	45	52	49	53	37	58	49	45	54
Weak lists									
List items									
Studied	67	64	73	66	62	73	66	63	73
Nonstudied	27	24	16	20	33	23	23	29	19
Difference	41	39	57	46	29	51	43	34	54
Critical items									
Studied	61	58	53	62	60	65	61	59	59
Nonstudied	17	20	13	23	23	15	20	22	14
Difference	44	38	39	39	37	50	41	37	45

Note. Confidence data are collapsed across "sure old" and "probably old" judgments and "remember"/"know" data are collapsed across "R" and "K" judgments. Data in the last three columns are collapsed across the two test formats.

obtained with analysis of the raw data. A 2 (item type) × 2 (list type) × 2 (test format) × 3 (presentation rate) ANOVA on these data indicated a main effect of list type, $F(1, 84) = 5.82$, $MSE = .063$, and a main effect of rate, $F(2, 84) = 3.29$, $MSE = .134$. No other main effects or interactions were significant. As in the previous experiment we computed separate ANOVAs on false recognition and true recognition, collapsing across test format because there were no main effects or interactions involving this variable. Data collapsed across test format can be found in the last three columns of Table 4.

Turning to the critical items first, a 2 (list type) × 3 (presentation rate) ANOVA revealed a significant main effect of list type, $F(1, 87) = 3.83$, $MSE = .076$, demonstrating that false recognition from strong lists (mean = .49) was greater than that from weak lists (.41), collapsing across rates. There was no main effect of rate and no interaction (both F's < 1). Although these list-level differences in false recognition are consistent with the previous experiment, the absence of an effect of rate on false recognition is not. Apparently, the effect of rate on false recognition in Experiment 2 was driven by prior recall. The same analysis on list items indicated a main effect of list type, $F(1, 87) = 4.28$, $MSE = .024$. This shows that true recognition from strong lists (.48) was greater than that from weak lists (.44), collapsing across rate. There was also a significant main effect of presentation rate, $F(2, 87) = 6.77$, $MSE = .062$, and this variable interacted with list type, $F(2, 87) = 3.43$, $MSE = .024$. The interaction suggests that rate had a larger effect on recognition of strong list items than weak list items, although follow-up analyses indicated that the effects were significant in each case.

Finally, list items and critical items from studied lists were recognized more often than their respective controls (i.e., baserate false alarms). Collapsing across rate and test format, this held for strong list items (.75 vs. .26, respectively, $t(89) = 21.63$, $SEM = .022$), weak list items (.67 vs. .24, $t(89) = 18.43$, $SEM = .024$), strong critical items (.82 vs. .34, $t(89) = 13.45$, $SEM = .037$), and weak critical items (.60 vs. .19, $t(89) = 13.92$, $SEM = .03$). These findings demonstrate robust true and false memory in all conditions, despite differences between lists in the overall magnitude of the effects.

Subjective judgments

We turn lastly to the subjective judgments (i.e., confidence and "remember"/"know") made to list items and critical items on the recognition test. As previously discussed, the pattern of results from each test format, when expressed simply as proportion of items called "old," were remarkably similar. This similarity was also upheld when we examined only those items that were recognized with high confidence (i.e., a "sure old" judgment) or with a "remember" judgment, both of which could be argued to reflect instances where subjective remembering was particularly compelling. Table 5 presents these data for each item type in each of the conditions, following the same format as Table 4. That is, the proportion of items recognized as "sure old" are listed for the confidence format, and the proportion of items recognized as "remembered" are listed for the "remember"/ "know" format. As can be seen, the pattern of results from the two testing formats was quite similar, and with only a few unsystematic exceptions, the means of comparable conditions across the two formats corresponded well with each other. Note that this similarity of results does not necessarily indicate that "remember" judgments are synonymous with high confidence judgments (for dissociations between the two see Gardiner & Java (1990) and Rajaram (1993)), but it does suggest that the same types of retrieved information can influence both judgments (cf. Tulving, 1985).

As with the overall proportions of "old" judgments, we collapsed across test format and analyzed adjusted false recognition and true recognition separately. A 2 (list type) × 3 (presentation rate) ANOVA on the proportion of critical items falsely recognized yielded a main effect of list type, $F(1, 87) = 28.26$, $MSE = .065$, with no effect of rate and no interaction (both p's $> .10$). The main effect of list type demonstrates that false recognition to critical items from strong lists was more compelling than that from weak lists. Collapsing across rate, separate analyses indicated this was true for "remember" judgments (mean = .42 for strong lists and .30 for weak lists, $t(44) = 2.09$, $SEM = .058$) as well as for high confidence judgments (.53 vs. .25, respectively, $t(44) = 6.18$, $SEM = .046$). A similar analysis on list items, again collapsing across test formats, indicated a main effect of rate, $F(2, 87) = 4.61$, $MSE = .063$, demonstrating that confidence and "remember" judgments increased with slower rates, as would be expected. There was also an effect of list type, $F(1, 87) = 6.58$, $MSE = .014$, and no interaction ($F < 1$). This effect suggests that recognition of strong list items was more compelling than that of weak list items. Separate analysis indicated that this difference was reliable for "remember" judgments (mean = .40 vs. .33, respectively, $t(44) = 2.65$, $SEM = .028$) but not for high confidence judgments (.46 vs. .44, $t < 1$).

As in the previous experiment, we analyzed unadjusted "remember" and high confidence judgments expressed as a proportion of the total probability of false alarms, collapsing across rate. With this comparison the difference in "remember" judgments to critical items from strong lists (.65) and weak lists (.59) failed to reach significance, $t(43) = 1.48$, $SEM = .041$, $p > .10$, although the effect was in the predicted direction. (One subject was excluded from this last analysis because they did not say "old" to any strong critical items.) With respect to high confidence false alarms, strong critical items (.75) yielded significantly more false recognition than weak critical items (.52), $t(44) = 5.21$, $SEM = .045$. A similar analysis on list items demonstrated no difference for either "remember" judgments (mean = .67 for strong lists and .64 for weak lists, $t < 1$) or high confidence judgments (.71 vs. .73, respectively, $t(44) = 1.04$, $SEM = .022$, $p > .10$). The finding that some critical items elicit more high confidence and "remember" false alarms than others suggests that false recognition from strong lists was more compelling than that from weak lists.

To summarize, the results of this experiment demonstrate that lists varied not only in the probability with which they elicited false recognition, but also in terms of how subjectively compelling the illusory memory was. False recognition from strong lists was accompanied with more "remember" and high confidence judgments than that from weak lists. Further, these differences were maintained when subjective judgments were adjusted for baserates, and also when they were expressed as proportions of the total probability of falsely recognized items in each condition (although this latter difference did not reach significance for "remember" judgments). Despite these differences across lists, it is worthwhile to note that "remember" and high confidence judgments to falsely recognized critical items were always greater when the list had been studied (Table 5, Rows 3 and 7) than when the list had not been studied (Rows 4 and 8). Thus, even weak lists elicited high levels of subjectively compelling false recognition relative to this nonstudied baserate.

Table 5
Mean percentage of list items (collapsed across serial positions 1, 8, and 10) and critical items recognized with high confidence or a "remember" judgment in Experiment 3

List type	Presentation rate								
	Confidence format			Remember/Know format			Mean		
	.5 s	1 s	3 s	.5 s	1 s	3 s	.5 s	1 s	3 s
Strong lists									
List items									
Studied	50	48	59	51	44	57	50	46	58
Nonstudied	9	3	7	12	12	7	11	8	7
Difference	41	45	52	39	32	50	40	38	51
Critical items									
Studied	62	71	54	62	53	51	62	62	52
Nonstudied	12	13	2	10	19	12	11	16	7
Difference	50	57	52	52	34	39	51	46	46
Weak lists									
List items									
Studied	48	46	57	42	39	48	45	42	53
Nonstudied	9	4	4	7	14	9	8	9	6
Difference	39	41	53	34	25	39	37	33	46
Critical Items									
Studied	33	28	27	44	33	33	38	30	30
Nonstudied	2	10	2	8	7	5	5	8	3
Difference	32	18	25	35	27	28	33	22	26

Note. Data in the last three columns are collapsed across the two test formats.

Comparing Experiments 2 and 3

Other than the lack of an effect of rate on false recognition, there were two notable differences between these recognition results and those of Experiment 2. To quantify these differences we directly compared the results of the two experiments, which was justified because the methodology between the two was similar by design, as was the subject population. The first difference is that list items were not recognized as often in the present experiment as they were in the previous experiment. Comparing unadjusted hit rates (collapsing across rate), this was true for both strong lists (mean = .75 and .84, respectively, $t(178) = 5.31$, $SEM = .017$) and weak lists (.67 vs. .81, $t(178) = 6.71$, $SEM = .02$). (Even larger effects were found when baserate-adjusted hit rates were compared, for reasons discussed below.) This testing effect is consistent with previous research that has demonstrated that the act of prior recall can boost recognition performance for list items (e.g., Roediger & McDermott, 1995). A similar effect may have operated on critical items, but the differential effects of rate across the two experiments make this difficult to determine.

A more dramatic difference between the two studies was that false alarms to control items from nonstudied lists (i.e., baserates) in the present experiment were considerably higher than they were in the previous experiment. Comparing Tables 3 and 4 it is clear that this was true for every item type in every condition. In fact, even the smallest difference between the two experiments was statically reliable (i.e., baserates for weak list items were .19 in the 3 s condition of the present experiment and .09 in that of Experiment 2, $t(58) = 3.17$, $SEM = .033$). We did not expect such differences between experiments, and we offer a potential explanation in the next section.

General discussion

The results of these experiments demonstrate an important point about the DRM false memory illusion: Lists that are constructed just like those of Roediger and McDermott (1995) (the first 15 associates of a nonstudied word) can fail to elicit high levels of false recall, despite this associative connection. The strong positive correlation between MBAS and false recall indicated that this is a critical factor in the degree to which lists will elicit such errors, as Deese (1959) demonstrated. We also extended Deese's original findings by demonstrating that false recognition was correlated with MBAS, consistent with McEvoy et al. (1999). An important new finding was that false recognition from lists with low MBAS was less compelling than that from lists with higher MBAS, as indexed by confidence ratings and "remember" judgments (Experiment 3). Another (rather surprising) finding was that list-level differences were also evident in false alarms to critical item controls (when the corresponding list had not been studied), as these baserates were higher for strong critical items than weak critical items. Finally, we demonstrated that slowing presentation rates decreased false recall and recognition from both strong and weak lists (Experiment 2), but the effects of rate on false recognition disappeared when prior recall was eliminated (Experiment 3). Therefore, when recognition is not compromised by prior recall, our results more closely resembled those of Arndt and Hirshman (1998).

In what follows we consider each of these results in turn, as well as their theoretical implications. The strong relationship between MBAS and false remembering constrains theories of how the information that leads to DRM false remembering is represented in memory. MBAS is an index of the degree of association between the list items and the critical item, and thus it can be considered a direct predictor of the amount of associative activation that the critical item receives when its associates are processed. Within the activation/monitoring theory, this activation could be conceived as automatic spreading activation within semantic memory that could leave a lasting effect, especially if it occurred repeatedly and/or yielded conscious thoughts of the critical item. Associates from lists with low MBAS would be less likely to activate the critical item, and therefore this information would not be as available during retrieval as it would be for lists with greater MBAS. As a result, subjects would be less likely to erroneously believe that weak critical items occurred in the study list relative to strong critical items. This activation/monitoring theory can account for the similar effects of associative strength on false recall and false recognition, because in both instances subjects would have falsely remembered the critical item to the degree that it was activated by list items and this activation was attributed to study (see Roediger, Balota, & Watson, 2001b,c).

The notion of differential activation from list to list can also explain why lists that elicited more false recognition also elicited more compelling false recognition, as indexed by confidence ratings and "remember" judgments. If lists with greater MBAS resulted in more activation of the critical item, then this additional activation may have made these items more retrievable (and/or more familiar) at test. Differences in familiarity could explain list-level variability in high confidence judgments, but the fact that "remember" judgments followed a similar pattern as high confidence judgments suggests that lists also varied in their ability to evoke false recollection of the nonpresented item's occurrence. Such false recollections could be explained by conscious activation (thoughts) of the critical item at study, which the subject later recollected and misattributed to actual presentation. If it is assumed that the probability of recollecting such thoughts depends on their prior frequency, and also that their prior frequency depends on the strength of association between the list and the critical item, then one might expect more "remember" judgments following more strongly related lists. Alternatively, processes other than the recollection of prior thoughts may cause "remember" judgments. For instance, in an effort to determine if a critical test item had been studied, the subject may imagine the critical item's study presentation upon encountering it at test. Depending on the level of familiarity of this item (due to associative activation), this imagination may be confused with actual recollection and result in a "remember" judgment (for further discussions of memory attributions, see Jacoby, Kelley, & Dywan, 1989; Johnson et al., 1993).

Gist or schema-based theories and semantic feature-based theories would have difficulty accounting for the relationship between MBAS and false remembering without making an additional critical assumption. Specifically, they would have to assume that strong associates share more semantic overlap with the critical item than do weak associates, and therefore a stronger gist trace (or more overlapping semantic features) is encoded for lists with greater MBAS. We are hesitant to make this assumption because the types of associations that are measured by free association tasks (and hence MBAS) are not necessarily semantic, or based on similarity of meaning. Instead, these associations are probably driven by a variety of factors (e.g., statistical co-occurrence in natural language) that could cause one concept to activate another. A more direct test of these theories would require a more direct measure of gist or semantic overlap for DRM lists, one that does not rely on associations, but we know of no such measure. Based on intuition, lists with strong MBAS do not appear to have more semantic overlap with the critical item than lists with weak MBAS (at least when forward associative strength is relatively constant). Indeed, in all of these lists there are many items that have semantic overlap with the critical item, as demonstrated by the *bitter-sweet* example used in the Introduction. Both lists appear to converge on the meaning of their critical items, but they differ greatly in MBAS (.011 and .172, respectively), and also in the mean probability of false recall and recognition.[5] This and other examples suggest that the semantic overlap account, although potentially applicable (see below), needs to be more rigorously defined in order to determine if it can account for the variability among lists to elicit false remembering.

We have argued that the variability among lists to elicit false remembering is most consistent with theories that propose that the critical item is activated from semantic memory and forms a lasting representation. Other evidence that such activation occurs in DRM is that the presentation of these lists yields priming of the critical item on implicit memory tests (McDermott, 1997; McKone & Murphy, 2000). The tests used in these studies are thought to be sensitive to both perceptual representations and the activation of abstract lexical representations (e.g., Rajaram & Roediger, 1993), suggesting that the lexical representation of the critical item was activated. It is unclear how gist-based representations, which by definition are not item-specific, could support such priming without additional assumptions. This is not to say that such mechanisms play no role in DRM false memories. Most of our lists that elicited the lowest levels of false recall still resulted in false recognition that was well above the false alarm rate to unrelated distracters. These

[5] A reviewer suggested that the *bitter* list converges on two different meanings of its critical item, whereas the *sleep* list does not, and that this might be the cause of the differences between these lists. However, close inspection of our other weak lists reveals that this is the exception, and not the rule. Further, it is unclear why activation of different meanings of the same word would be predicted to yield low levels of false remembering. The *smell* list consists of words that converge on the verb (*breathe*, *sniff*, *hear*, *see*, *whiff*) the noun (*aroma*, *scent*, *reek*, *stench*, *fragrance*) and that are in other ways associated to *smell*, and this list elicited one of the highest levels of false recall (60%) in the Stadler et al. (1999) norms.

relatively high false recognition rates make sense if one assumes that familiarity (via semantic feature or gist overlap) may contribute more to recognition tests (where the test item is presented to the subject for a decision) than recall tests (see Mandler, 1980; Jacoby, 1991).

Perhaps the strongest evidence that feature-based or gist-based processes are involved, in addition to the activation processes advocated here, comes from the differential effects of delay on true and false memories. One prediction of prototype-based models is that abstract representations should be more resistant to forgetting than memory for individual exemplars (e.g., Posner & Keele, 1970). Similarly, fuzzy trace theory posits that gist traces are more resistant to forgetting (or interference) than verbatim (or item-specific) traces (e.g., Reyna & Lloyd, 1997), and exemplar-based or feature-based models also predict greater acceptance of prototypes than exemplars after a delay (e.g., Hintzman, 1986). Consistent with this prediction, several researchers have found that recall of list items decreases more over a delay than recall of critical items (e.g., Seamon et al., in press a; Toglia et al., 1999), although the picture is less clear for true and false recognition (e.g., Lampinen & Schwartz, 2000; Seamon et al., in press a; Thapar & McDermott, 2001). Activation-based theories have difficulty accounting for this interaction between item type and retention interval, especially when rates of true and false remembering were equivalent on the first memory assessment.

In addition to these processes that characterize the nature of the information that causes false remembering, the fact that slower presentation rates decreased false recall in Experiment 2 suggests that additional processes are involved. As mentioned earlier, one candidate process is the monitoring of memory accuracy (see also McDermott & Watson, 2001). Slower presentation rates may have provided more time to engage in distinctive item-specific processing of the lists, such as mental imagery or rehearsal. As a result, even though the critical item may have been equally activated/encoded at every rate, or even more activated/encoded at slower rates, this would have been offset by subjects' increased utilization of item-specific recollections to distinguish between presented and nonpresented words. For instance, at slower rates subjects who thought of the critical item may have been more likely to realize (during study) that it was not presented, thereby reducing subsequent source confusions (see Bredart, 2000). Monitoring processes may also have occurred at test. At slower rates subjects may have demanded more distinctive item-specific recollections before producing an item on the recall test. As a result, subjects would have been less likely to recall nonpresented items that did not benefit from enhanced item-specific processing at slower study rates.

Gist-based theories could explain the effect of presentation rate by assuming that slowing rate differentially increases the storage of verbatim traces, relative to gist traces. As a result, subjects might have relied more on verbatim than gist traces at retrieval, which would have reduced false remembering. Exactly how subjects would switch reliance on these different types of traces is unknown, but one explanation is that they use monitoring processes similar to those discussed above. It is less clear how semantic feature-based models would account for the present effects of rate on false recall, because they predict increases in false remembering under conditions where feature encoding should be enhanced (i.e., slower rates). However, if a monitoring process were applied to the output of these models then they too could fit the obtained pattern.

One unexpected result was that slowing rate decreased false recognition when the test was preceded by recall (Experiment 2), but not when it was preceded by math (Experiment 3). This finding is inconsistent with a monitoring account, because the monitoring process should presumably operate during either recall or recognition tests. One explanation is that in the present experiments recall always came immediately after a list's presentation, where monitoring processes that rely on item-specific memory should be greatest, and recognition tests were administered after all lists had been studied. This may have made the recall tests more sensitive to monitoring processes than the recognition tests, leading to an effect of rate on the former but not the latter. This may not matter, though, because Arndt and Hirshman (1998) also found that slowing rate did not reduce false recognition, and their results were based on recognition tests administered after each list's presentation. A more intriguing possibility is that recall is, in general, more sensitive to such monitoring processes than recognition (see Koriat & Goldsmith, 1996). Thus, it may require more dramatic changes in presentation rate to affect false recognition than false recall.

Another question that arises from the previous analysis is that, if slowing presentation rate was insufficient to support effective monitoring on our recognition tests, what drove the effects of rate on false recognition in Experiment 2? As mentioned earlier, the major methodological difference between Experiments 2 and 3 was that all lists were recalled prior to recognition in Experiment 2. Prior recall probably affected false recognition in two ways. The first is that, in the fast rate conditions, subjects were more likely to recall critical items, and were thus more likely to later falsely recognize these items due to a failure of source-monitoring. However, this factor alone is not sufficient to account for the present results. If the effects of rate on false recognition in Experiment 2 were due solely to prior recall of critical items, then a conditional analyses should reveal no effects of rate on false recognition for those critical items that were not previously recalled. This was not the case. For strong lists, the conditionalized probability of false recognition was .70 at the .5 s rate, .63 at the 1 s rate, and .47 at the 3 s rate, $F(2, 89) = 5.12$, $MSE = .078$. For weak lists these probabilities were .36 at the .5 s rate, .39 at the 1 s rate, and .23 at the 3 s rate, $F(2, 89) = 4.36$, $MSE = .047$. Such conditional analyses should be interpreted with caution, as they are susceptible to item-selection artifacts, but such artifacts cannot explain the results for weak lists, because these items were rarely falsely recalled (and hence conditionalized scores are not much different than the unconditionalized scores). Because rate affected false recognition even for those items that were not previously recalled, some factor other than previous recall of critical items was involved.

A likely explanation is that prior recall boosted memory for list items, which is reflected in the greater recognition of these items in Experiment 2 than Experiment 3. Because more list items were recalled at slower presentation rates, one would expect this testing effect to be greatest in the 3 s condition. One result of such a testing effect would be to enhance monitoring strategies that rely on recollection of list items. Thus, although the recognition test may not have been sensitive to monitoring processes supported by the effects of rate alone (Experiment 3), these effects of rate, in addition to the effects of prior recall on list items, may have enhanced the effectiveness of monitoring processes in Experiment 2. It is less clear how gist-based theories would explain this effect of prior recall; according to fuzzy-trace theory prior recall in the DRM paradigm should strengthen gist traces and hence increase false recognition (Brainerd et al., 2001). Thus, if anything, increasing prior recall should have led to more false recognition as rate decreased, but the opposite result was found. The notion that prior recall enhanced monitoring processes also explains the otherwise puzzling finding that false alarms to list items and critical items from nonstudied lists (i.e., control items) were lower in Experiment 2 than in Experiment 3, across the range of presentation rates. The enhanced recollection of list items due to prior recall may have facilitated the discrimination between list items and unrelated lures, leading to fewer false alarms of these items in Experiment 2.

We turn lastly to false alarms to control items. As discussed, previous research has found that these baserates are somewhat greater for critical items than for list items, and in the present experiments we found them to be greater for strong critical items than for weak critical items. These baserate differences did not compromise our major findings, because our results were maintained even after these baserates were taken into account (via subtraction). Nevertheless, these baserates are interesting to consider in their own right. One possibility is that these control items received some associative activation from related items that happened to occur before them on the test. Marsh, McDermott, and Roediger (in press) reported that critical item controls were more likely to be falsely recognized if they were preceded by related list items on the test (but this manipulation did not affect critical items from studied lists). Because strong critical items are more highly related to their list associates than are weak critical items, differences in such test activation could have driven the baserate differences found here. Regardless of the explanation, the fact that such differences do exist suggests that results from studies in which background false alarm rates are not experimentally measured should be interpreted with caution.

Considered as a whole, the evidence from these new experiments (as well as other results reviewed here) demonstrates that the creation of false memories in the DRM paradigm involves at least two factors: (1) the activation and/or encoding of information that can cause false remembering and (2) monitoring or editing processes that modulate the extent to which this information yields false remembering. The strong relationship between MBAS and false recall and false recognition is most consistent with theories that emphasize associative activation of the critical item, as opposed to gist or semantic feature-based theories (see Roediger et al., 2001b). These other mechanisms are probably involved, with the relative contribution of each type of information potentially depending on task dynamics (e.g., recall vs. recognition, retention interval). More importantly, the wide range of variability among lists to elicit this illusion indicates that it is highly sensitive to subtle stimulus manipulations, whereas the effects of rate on false recall suggest that the illusion is also sensitive to monitoring processes. These two observations indicate that both the stimuli and the individual exert control over false memories. However, it should be stressed that most of the DRM demonstrations of individual or cognitive control depend on monitoring processes that, in turn, depend on stimulus characteristics (e.g., presentation rate, presentation format, presentation modality, and number of presentations). As previously discussed, even instructing subjects to avoid this illusion via warnings and examples (arguably the loudest call to arms of cognitive control) was most effective when given before study, when it could influence how the stimuli were initially processed. Based on these findings, it appears that individual control of memory accuracy primarily depends on characteristics of the to-be-remembered material.

Acknowledgments

Experiments 1 and 2 of this research were included in a Masters thesis submitted to Washington University by D.G., and a talk delivered by H.R. at the 39th Annual Meeting of the Psychonomic Society in Dallas (November, 1998). We thank Dave Balota and Mitch Sommers for helpful discussions, and Charles Brainerd, Tim Curran, and John Seamon for thoughtful comments. We also thank Carrie English and Angela Wong for assistance collecting data.

Appendix A

The 28 lists used in Experiment 1, arranged alphabetically by critical item.

Beautiful: ugly, pretty, girls, woman, homely, lovely, nice, picture, lady, mountain, snow, scene, music, day, gorgeous.

Bitter: sweet, sour, taste, chocolate, rice, cold, lemon, angry, hard, mad, acid, almonds, herbs, grape, fruit.

Butterfly: moth, insect, wing, bird, fly, yellow, net, flower, bug, cocoon, summer, color, bee, stomach, worm.

Cabbage: head, lettuce, vegetable, food, salad, green, garden, leaf, sauerkraut, smell, slaw, patch, plant, carrots, soup.

Carpet: rug, floor, soft, red, sweeper, tack, walk, bag, room, blue, chair, thick, deep, magic, wool.

Citizen: United States, man, person, American, country, alien, people, vote, me, patriot, flag, foreigner, France, immigrant, member.

City: town, crowded, state, capital, streets, subway, square, New York, village, metropolis, big, Chicago, suburb, county, urban.

Command: order, army, obey, officer, performance, do, tell, general, shout, halt, voice, soldier, harsh, attention, sharp.

Cottage: house, lake, cheese, home, white, cabin, small, door, fence, vines, woods, ivy, roses, cozy, hut.

Foot: shoe, hand, toe, kick, sandals, soccer, yard, step, ankle, arm, boot, inch, sock, knee, mouth.

Health: sickness, good, happiness, wealth, ill, doctor, service, strong, hospital, disease, body, vigor, center, pain, robust.

Justice: peace, law, courts, judge, right, liberty, government, jury, truth, blind, fair, supreme, crime, department, trial.

King: queen, England, crown, prince, George, dictator, palace, throne, chess, rule, subjects, monarch, royal, leader, reign.

Lamp: light, shade, table, bulb, post, black, cord, desk, bright, lighter, read, on, bed, burn, stand.

Long: short, fellow, narrow, John, time, far, hair, island, road, thin, underwear, distance, line, low, rope.

Mutton: lamb, sheep, meat, chops, beef, veal, collar, leg, eat, fat, coat, stew, fur, pork, steak.

Needle: thread, pin, eye, sewing, hole, point, prick, thimble, haystack, thorn, hurt, injection, syringe, cloth, knitting.

Pen: pencil, write, fountain, leak, quill, felt, Bic, scribble, crayon, Cross, tip, marker, ink, cap, letter.

Rough: smooth, bumpy, sea, tough, sandpaper, jagged, ready, coarse, uneven, riders, rugged, sand, boards, ground, gravel.

Rubber: elastic, bounce, gloves, tire, ball, eraser, springy, foam, galoshes, soles, latex, glue, flexible, resilient, stretch.

Smoke: cigarette, puff, blaze, billows, pollution, ashes, cigar, chimney, cough, tobacco, stink, match, lungs, flames, stain.

Stove: hot, heat, pipe, cook, warm, fire, oven, wood, kitchen, lid, coal, gas, iron, range, furnace.

Swift: fast, slow, river, Jonathan, current, rapid, stream, water, quick, Gulliver, run, sure, deer, car, author.

Trash: garbage, waste, can, refuse, sewage, dirt, junk, rubbish, paper, scraps, pile, dump, landfill, debris, litter.

Trouble: bad, shooter, worry, danger, sorrow, fear, school, problem, police, fight, sad, difficulty, help, maker, jail.

Whiskey: drink, drunk, beer, liquor, gin, bottles, alcohol, rye, glass, wine, rum, bourbon, evil, bar, scotch.

Whistle: stop, train, noise, sing, blow, tune, sound, dog, song, shrill, boy, lips, wolf, call, loud.

Wish: want, dream, desire, hope, well, think, star, bone, ring, wash, thought, get, true, for, money.

References

Anastasi, J. S., Rhodes, M. G., & Burns, M. C. (2000). Distinguishing between memory illusions and actual memories using phenomenological measurements and explicit warnings. *American Journal of Psychology, 113*, 1–26.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology, 77*, 171–179.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: explanations from a global matching perspective. *Journal of Memory and Language, 39*, 371–391.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 941–947.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*, 160–179.

Brainerd, C. J., Wright, R., Reyna, V. F., &Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 307–327.

Bredart, S. (2000). When false memories do not occur: not thinking of the lure or remembering that it was not heard? *Memory, 8*, 123–128.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22.

Gallo, D. A., McDermott, K. B., Percer, J. M., & Roediger, H. L. (2001a). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 339–353.

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: can we avoid creating false memories? *Psychonomic Bulletin & Review, 4*, 271–276.

Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001b). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8*, 579–586.

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

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18*, 23–30.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8–20.

Hicks, J. L., & Marsh, R. L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1195–1209.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Pscyhological Review, 95*, 528–551.

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

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4*, 577–581.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67–85.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lampinen, J. M., & Schwartz, R. M. (2000). The impersistence of false memory persistence. *Memory, 8*, 393–400.

Libby, L. K., & Neisser, U. (2001). Structure and strategy in the associative false memory paradigm. *Memory, 9*, 145–163.

Loftus, E. F., & Palmer, J. C. (1974). Recon truction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585–589.

Mandler, G. (1980). Recognizing: the judgment of previous occurrence. *Psychological Review, 87*, 252–271.

Marsh, E. J., McDermott, K. B., & Roediger, H. L. (in press). *Does test-induced priming play a role in the creation of false memories*? *Memory*.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language, 35*, 212–230.

McDermott, K. B. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. *Psychonomic Bulletin & Review, 4*, 582–586.

McDermott, K. B., & Roediger, H. L. (1998). Attempting to avoid illusory memories: robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language, 39*, 508–520.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: the impact of presentation duration. *Journal of Memory and Language, 45*, 160–176.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: a time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 563–582.

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What is the connection between true and false memories? The differential roles of interitem associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1177–1194.

McKone, E., & Murphy, B. (2000). Implicit false memory: effects of modality and multiple study presentations on long-lived semantic priming. *Journal of Memory and Language, 43*, 89–109.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: the role of criterion shift in false memory. *Psychological Review, 106*, 398–405.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). *The University of South Florida word association, rhyme, and word fragment norms*, University of South Florida, Tampa (unpublished).

Neuschatz, J. S., Payne, D. G., Lampinen, J. M., & Toglia, M. P. (2001). Assessing the effectiveness of warnings and the phenomenological characteristics of false memories. *Memory, 9*, 53–71.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35*, 261–285.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.

Rajaram, S. (1993). Remembering and knowing: two means of access to the personal past. *Memory & Cognition, 21*, 89–102.

Rajaram, S., & Roediger, H. L. (1993). Direct comparison of four implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 765–776.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy trace theory: an interim synthesis. *Learning and Individual Differences, 7*, 1–75.

Reyna, V. F., & Lloyd, F. (1997). Theories of false memory in children and adults. *Learning and Individual Differences, 9*, 95–123.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science, 8*, 231–237.

Roediger, H. L., Balota, D. A., & Robinson, K. J. (2001a). *Automatic mechanisms in the creation of false memories*. Washington University, St. Louis (unpublished).

Roediger, H. L., Balota, D. A., & Watson, J. M. (2001b). Spreading activation and the arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington DC: American Psychological Association Press.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Roediger, H. L., & McDermott, K. B. (1999). False alarms and false memories. *Psychological Review, 106*, 406–410.

Roediger, H. L., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. In M. A. Conway, S. E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory II* (pp. 187–245). Hove, Sussex: Psychological Press.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001c). Factors that determine false recall: a multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: evidence from ROC curves. *Journal of Memory and Language, 43*, 67–88.

Russell, W. A., & Jenkins, J. J. (1954). *The complete Minnesota norms for responses to 100 words from the Kent–Rosanoff Word Association Test*. Technical Report No. 11, Contract N8 ONR 66216, Office of Naval Research, University of Minnesota, MI.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: the distinctiveness heuristic. *Journal of Memory and Language, 40*, 1–24.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: false recall and recognition in amnesic patients. *Journal of Memory and Language, 35*, 319–334.

Seamon, J. G., Luo, C. R., & Gallo, D. A. (1998). Creating false memories of words with or without recognition of list items: evidence for nonconscious processes. *Psychological Science, 9*, 20–26.

Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (in press a). Are false memories more difficult to forget than accurate memories?: the effect of retention interval on recall and recognition. *Memory & Cognition*.

Seamon, J. G., Luo, C. R., Schwartz, M. A., Jones, K. J., Lee, D. M., & Jones, S. J. (2002). Repetition can have similar or different effects on accurate and false recognition. *Journal of Memory and Language, 46*, 323–340.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*, 710–715.

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*, 494–500.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: effects of level of processing and retention interval. *Memory & Cognition, 29*, 424–432.

Toglia, M. P., & Neuschatz, J. S. (1996, November). *False memories: Where does encoding opportunity fit into the equation*? Poster session presented at the annual meeting of the Psychonomic Society, Chicago.

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: when more is less. *Memory, 7*, 233–256.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26*, 1–12.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70*, 122–129.

Watkins, M. J. (1991). An experimental psychologist's view of cognitive science. In R. G. Lister & H. J. Weingartner (Eds.), *Perspectives on cognitive neuroscience* (pp. 132–144). New York: Oxford University Press.

Whittlesea, B. W. A., & Leboe, J. (2000). The heuristic basis of remembering and classification: fluency, generation, and resemblance. *Journal of Experimental Psychology: General, 129*, 84–106.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical design theory: comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review, 107*, 377–383.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*, 368–376.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: the contribution of recollection and familiarity. *Memory & Cognition, 25*, 747–763.