Psychological Review
1999, Vol. 106, No. 2, 398–405

0033-295X/99/$3.00

Theoretical Commentary: The Role of Criterion Shift in False Memory

Michael B. Miller and George L. Wolford
Dartmouth College

H. L. Roediger and K. B. McDermott (1995) reintroduced a paradigm originally developed by Deese (1959). According to the authors, the paradigm provides a technique for the creation of false memories. The paradigm is reliable and easy to implement. Because of these characteristics and the current interest in false memories, the paradigm has been used in many recent studies. The authors replicated Roediger and McDermott's results in two experiments. When conditions were included that allowed the computation of signal-detection parameters, it was found that most of the false memories could be ascribed to criterion shifts. The authors discuss the possible role of criteria in defining and understanding false memories.

Memory distortions have been of interest to researchers for some time. Formal demonstrations of such distortions date at least to the work of Bartlett (1932). Interest in those distortions has been much heightened by research on eyewitness testimony (Ceci & Bruck, 1993; Loftus, 1979) and recovered memories (Lindsay & Read, 1994; Loftus, 1993). Roediger and McDermott (1995) reintroduced a paradigm for studying false memories originally developed by Deese (1959). Because the paradigm is reliable and easy to use, it has been adopted by a number of researchers interested in studying the characteristics of false memories (Johnson et al., 1997; McDermott, 1996; Miller & Gazzaniga, 1998; Payne, Elie, Blackwell, & Neutschatz, 1996; Read, 1996; Schacter, Reiman, et al., 1996; Schacter, Verfaellie, & Pradere, 1996).

The paradigm reintroduced by Roediger and McDermott (1995) consists of presenting participants with lists of words to memorize. The words are chosen for their high association to a single target word, such as *sleep,* but the word *sleep* is not presented. These target words are called *critical lures.* The nonpresented critical lures are recalled during free recall at well-above-chance frequencies, even though intrusions are usually rare in free recall. Even more impressively, the critical lures are labeled "old" by the participants on a subsequent recognition test approximately as often as words that were actually presented. The high percentage of "old" responses to the critical lures is taken as evidence that false memories were created.

In describing recognition performance more generally, many researchers have argued that the percentage of "old" responses to a particular item type is some combination of memory for the items (sensitivity) and decisions about how liberal or conservative to be in judging an item as old (criterion; e.g., Murdock, 1974; Ratcliff, Sheu, & Gronlund, 1992). Presumably the high percentage of old responses to the critical lures could represent sensitivity (strengthening of the nonpresented critical lures as a result of list context), a lower criterion for calling critical lures old (based on the list context), or some combination of the two. The theory of signal detection (Green & Swets, 1966/1974) was designed to provide independent estimates of sensitivity and criterion.

We expanded the Roediger and McDermott (1995) paradigm to include the conditions required to perform a signal-detection analysis. A signal-detection analysis requires the percentage of old responses to each item type, both when it was presented and when it was not presented (hits and false alarms). The Roediger and McDermott paradigm includes three item types: critical lures, items related to the critical lures, and unrelated items. In the Roediger and McDermott paradigm, only related items were ever presented, so we added lists in which the critical lures were included and lists of unrelated items.

Another way to think about critical lures and false memories is to assume that because participants respond "old" as often to the critical lures as they do related items that were presented, it is as if the critical lures had been presented. Because they were not presented, the memories are false. However, we do not know what performance would be like on the critical lures if they had been presented, because that condition was not included. Performance on presented critical lures might be extremely high relative to either nonpresented critical lures or presented related items. That outcome would tend to indicate that there was a criterion shift rather than a strengthening or storage of the nonpresented critical lures. The critical lures differ from the other items both structurally and relationally. By definition, they are words with many high associates, and most of the high associates are presented. It is plausible to think that participants could set a different criterion to such words appearing on a recognition test. If this were true, then the high frequency of "old" responses to nonpresented critical lures could be the result of criterion shifts.

We replicated Experiment 2 of Roediger and McDermott (1995) as closely as possible in two experiments, except that we included additional conditions so that each of the item types could be

Michael B. Miller and George L. Wolford, Program in Cognitive Neuroscience, Dartmouth College.

The order of authorship was determined alphabetically. We thank Paul Corballis, Margaret Funnell, Michael Gazzaniga, Kevin Sailor, and Thomas Wickens for their assistance and support.

Correspondence concerning this article should be addressed to Michael B. Miller, Program in Cognitive Neuroscience, Dartmouth College, 6162 Silsby Hall, Hanover, New Hampshire 03755. Electronic mail may be sent to michael.b.miller@dartmouth.edu.

compared in a presented versus a nonpresented format. We presented lists with only related items (as in Roediger and McDermott), lists in which the critical lure displaced one of the related items, and lists of unrelated items.

Experiment 1

The two experiments that we carried out were identical except that participants were asked to make remember–know judgments in the first experiment and confidence judgments in the second experiment. The remember–know judgments were used by Roediger and McDermott in their second experiment.

Method

Participants. Twenty-four Dartmouth College undergraduates participated in the experiment. Participants were given extra credit toward their grade in an introduction to psychology course for their participation. Participants were tested in six groups ranging from 3 to 5 participants per group.

Materials. We used the same 24 lists from the Roediger and McDermott (1995) study. Each list contained 15 words (e.g., bed, rest, awake, tired, dream, wake, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy) that are all close associates of a target word, referred to as a critical lure (e.g., sleep). The order of the words in the list was held constant; the strongest associate appeared first, and words appeared in descending order by strength of association to the target word. In addition, we used 12 words from the lists that were not used as related lists and 18 new words unrelated to any of the lists (e.g., dog, tool, dollar, church) to make up two lists of unrelated words.

Design. We used a within-subjects design. Each participant received three types of lists: four lists containing only the 15 high associates to the critical lure as in Roediger and McDermott (1995), eight lists in which one of the related items was replaced by the critical lure, and two lists of unrelated items. During recognition, the participants were presented with some words from each of the three item types (critical lures, related items, and unrelated items) that had been presented and some from each item type that had not been presented. The presentation of the lists was rotated through the separate groups so that each word was tested as presented for some of the participants and as not presented for the other participants.

The 24 original lists of words were arbitrarily divided into two sets. Each participant heard one set, or 12 of the original lists, plus two lists of words unrelated to each other for a total of 14 lists. Three separate groups were tested for each set of lists, yielding a total of six groups. In four of the lists, the critical lure (e.g., sleep) was placed in the first position of the list, displacing the word originally held in that position (e.g., bed), the highest associate. The displaced word was not presented, but was used as a nonpresented related list item during the recognition test. In another four of the lists, the critical lure displaced the word originally held in the 10th position (e.g., slumber); again that 10th word was not presented and was used as a nonpresented related list item for testing. The positions were determined by replicating the position of the studied items used as test items in the original Roediger and McDermott (1995) study. In four lists, the critical lure was not presented.

The two lists of unrelated words were always presented in the middle position; six thematic lists were presented before and six thematic lists presented after the unrelated lists. The two unrelated lists consisted of six words from thematic lists that were not presented to the participant, including two words used as critical lures, two words from the first position of the lists, and two words from the 10th position of the lists. The other nine words making up the list were chosen from categories unrelated to the thematic lists, matched for frequency and concreteness. These latter nine words were not used as recognition test items. The six words from the unused thematic lists were placed into Positions 1, 2, 6, 8, 10, and 15 of the study lists based on the averaged free-recall performance of those positions from a pilot study and matched to the averaged recall performance from Positions 1 and 10 used for the related list items.

Lists and words were counterbalanced across participants. Therefore, each critical lure appeared as a presented item in the first and 10th positions and as a nonpresented item. Related and unrelated items also appeared as presented and nonpresented. Participants studied the 14 lists of words, with half the lists (6 thematic lists and 1 unrelated list) immediately followed by a free-recall test and the other half followed by math problems. After the presentation of all the lists, a recognition test was given, with the participants responding either "old" or "new;" if they responded "old," they were asked to make a further remember–know judgment.

Procedure. The design, and particularly the procedure, closely followed the Roediger and McDermott (1995) study. The whole procedure took approximately 1 hr per group. Participants were told that they were being tested on their memory for words. The words were read by the experimenter at the rate of 1.5 s per word. After the presentation of each list, the participant would hear the experimenter say either the letter *A* or the letter *B*. Half of the participants in each group were assigned the letter *A* and the other half were assigned *B*. If the experimenter said *A*, the *A* participants were to recall as many words as they could remember from the list by writing them down, while the *B* participants were to work the math problems. If the experimenter said *B*, then the reverse was true. Participants were given 2 min to complete the free-recall or math problems, at which point they would hear a tone and the experimenter saying "next list." Five seconds later, the experimenter presented the next list of words. Participants wrote down their responses on 1 of the 14 4- × 11-in. slips of paper with math problems at the top of the paper. They were instructed to turn over the slip of paper before the presentation of the next list.

Immediately after this portion of the testing, participants were given a recognition test. They were given brief verbal instructions on how to make old–new distinctions and remember–know judgments, and they read more detailed written instructions. They then were given a sheet with a list of 72 words, with the words "old" and "new" and a blank line next to each of them. Participants were to circle either "old" or "new" in response to each item; if they circled "old," they were to write either *R* or *K* in the blank. *R* is supposed to represent the conscious recollection of the word from the list, whereas *K* is supposed to represent recognition of the word without conscious recollection of its occurrence. The detailed instructions for making the *R-K* judgment were modeled after Rajaram (1993). The written instructions urged the participants to make the *R-K* judgment only on the basis of what they heard in the presentation of the study lists.

The recognition test comprised 72 items, 36 presented and 36 nonpresented. The items included 8 critical lures presented (in the first or 10th position), 4 critical lures not presented, 16 related list items presented (in the first or 10th positions), 8 related list items not presented (displaced from the first or 10th positions), 12 unrelated items presented (4 critical lures and 8 related list items each from the first and 10th position of nonpresented lists), and 24 unrelated items not presented (including 8 critical lures and 16 related list items each from the first and 10th positions of nonpresented lists). All the test items were randomly intermixed. Each group was given a new random order of the test items, but all participants within a group received the same test sheet. After the recognition test, participants were debriefed.

Results

We first examined performance on the critical lures relative to the other item types. Those results are summarized in Table 1 and Figure 1. The probability of saying "old" to a critical lure that was not presented was not significantly different from the probability of saying "old" to a related item that was presented (.81 vs. .88), $t(23) = 1.17, p = .254$. The probability of saying "old" to a critical lure that was not presented was very much higher than the prob-

Table 1
Performance on Recognition as a Function of Item Type

Item type	Experiment 1		Experiment 2
	P(old)	*P*(remember)	*P*(old)
Critical lures			
Presented	.97	.80	.96
Nonpresented	.81	.44	.78
Related items			
Presented	.88	.72	.86
Nonpresented	.36	.14	.42
Unrelated items			
Presented	.60	.44	.67
Nonpresented	.11	.03	.22

ability of responding "old" to an unrelated item that was not presented (.81 vs. .11), $t(23) = 11.11$, $p < .01$. Those values provide a similar pattern to that found in Roediger and McDermott (1995). One might draw the conclusion from Roediger and McDermott based on the similarity of nonpresented critical lures to presented related items that critical lures behave as if they had been presented. That conclusion is drawn into question by the fact that the actual presentation of the critical lure raises the probability of responding "old" to .97. The difference between presenting and not presenting the critical lure is highly significant, $t(23) = 3.21$, $p = .004$. In other words, performance on the nonpresented critical lures is not the same as if they had been presented.

Further evidence for the difference between the presented and nonpresented critical lures is contained in the remember–know judgments. When the critical lures were presented, 80% of the items were judged as remembered versus 17% judged as known. When the critical lures were not presented, only 44% of the items were judged as remembered versus 37% judged as known. That difference is highly significant, $t(23) = 6.34$, $p < .01$. So even when they are both judged "old," the critical lures that were presented are distinct from those that were not presented.

The final evidence for the difference between the presented and nonpresented critical lures is found in the free-recall data (see Figure 1). The critical lures that were not presented were recalled on the free-recall test 27% of the time. Five percent of all the responses in the free-recall were intrusions, of which 15% were critical lures. However, when the critical lures were actually presented, the probability of recall was 96% if presented in Position 1 and 83% if presented in Position 10.

The experiment was designed specifically to afford a signal-detection analysis. Table 2 contains measures of sensitivity and criterion for each of the three item types. We calculated a measure of sensitivity, $d(a)$, as proposed by Simpson and Fitter (1973). According to Macmillan and Creelman (1991), $d(a)$ is a particularly appropriate measure of sensitivity in a situation such as ours. The normal d' requires the assumption of equal variance in the presented and nonpresented distributions, but $d(a)$ allows for unequal variances and is preferred in single-point receiver operator characteristic (ROC) spaces. The measure of criterion (c2) is one proposed by Macmillan and Creelman (1991) for use when the variances of the two distributions are not equal. It reduces to the usual measure of criterion when the variances are equal. The measures of sensitivity were quite similar across the three item types, but the measures of criterion were quite different. Figure 2 provides a further illustration of this point. The hits are plotted against the false alarms for each of the item types on normal–normal coordinates. If the three points were on the same ROC curve (i.e., had the same sensitivity), they would lie on a straight line in normal–normal coordinates. The best fit straight line is shown in the figure, and that line accounts for 98% of the variance in the three points. The slope of the line was less than 1 (0.76), indicating that the variance of the nonpresented items is smaller than the variance of the presented items. Ratcliff et al. (1992) also

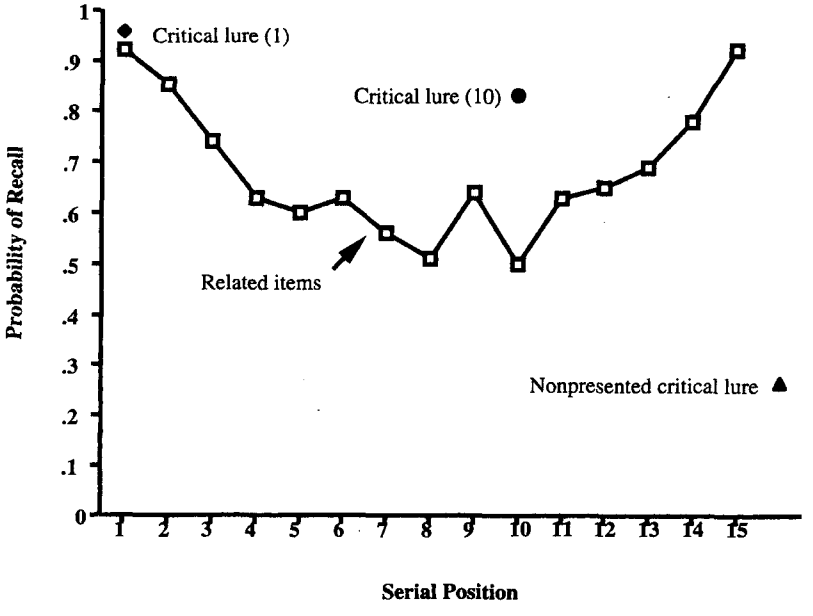

Figure 1. Probability of correct recall as a function of serial position in Experiment 1.

Table 2
Measures of Sensitivity and Bias as a Function of Item Type

	Experiment 1		Experiment 2		
Item type	$d(a)$	$c2$	$d(a)$	$d'(e)$	$c2$
Critical lures	1.37	−1.19	1.08	1.08	−1.19
Related items	1.63	−0.35	1.46	1.48	−0.41
Unrelated items	1.34	0.42	1.12	1.16	0.06

plotted ROC curves from several old–new recognition experiments on normal–normal coordinates and obtained slopes averaging 0.78 across experiments. The inequality of variances supports the use of $d(a)$ and c2 as the appropriate measures of sensitivity and criterion.

The item types differed substantially in criterion. The criterion c2 is calculated in standard score units, so the differences between item types is large. For instance, the criterion for critical lures was over one and one-half standard units lower than the criterion for unrelated items. If the criteria were equal, they would lie on a line more or less perpendicular to the best fit line in Figure 2, a situation that was patently not true.

In our view, the close similarity of the $d(a)$ values across item types coupled with the large differences in criterion argues strongly that the high false-positive rate with critical lures is the result of a shift to a lax criterion for those items. The equality of sensitivity across item types indicates that the critical lures profit as much as any other item type by being presented. Conversely, the critical lures are disadvantaged by not being presented as much as any other item type. The similarity of the values of false alarms on critical lures to hits on related items does not mean that equivalent memories were created. It means that critical lures yielded a lower criterion than related items.

Experiment 2

The second experiment was identical to the first one except that the participants were asked to make confidence ratings rather than remember–know judgments. We switched to confidence ratings because they allow the computation of more stable signal-detection parameters. Roediger and McDermott (1995) used confidence ratings in their first experiment but not for the computation of signal-detection parameters. The second experiment also affords a secondary analysis that was not available in Experiment 1. That analysis is described in the Discussion.

Method

Participants. Nineteen Dartmouth College undergraduates participated in the experiment. Participants were given extra credit toward their grade in an introduction to psychology course for their participation. Participants were tested in six groups ranging from 3 to 5 participants per group.

Materials. The same materials in Experiment 1 were used for this experiment.

Procedure. The only change in the procedure from Experiment 1 is in the design of the recognition test. In Experiment 1, participants were asked to make a remember–know judgment about each of the items judged to be old. In the recognition phase of the second experiment, participants were given the same test words on a sheet as in Experiment 1, but this time they were to make a single judgment on a 6-point scale. They were instructed to write next to each word a number from 1 to 6 corresponding to their confidence that the word was presented on one of the study lists (1 = *definitely old,* and 6 = *definitely new*). The scale was explained and printed on the test sheet for reference.

Results

We first examine performance on the critical lures relative to the other item types. Those results are summarized in Table 1 and Figure 3. The probability of saying "old" to a critical lure that was not presented was not significantly different from the probability of responding "old" to a related item that was presented (.78 vs. .86), $t(18) = 1.75$, $p = .098$. The probability of saying "old" to a critical lure that was not presented was very much higher than the probability of responding "old" to an unrelated item that was not presented (.78 vs. .22), $t(18) = 10.13$, $p < .01$. Presented critical lures are significantly more likely to be labeled "old" than nonpresented critical lures (.96 vs. .78), $t(18) = 3.78$, $p < .01$. Again, performance on the nonpresented critical lures is not the same as if they had been presented.

The free-recall data in Experiment 2 (see Figure 3) are similar to the free-recall data from Experiment 1. The critical lures that were not presented were recalled on the free-recall test 42% of the time (somewhat higher than in Experiment 1). Of all the responses in the free-recall test, 4.5% were intrusions, of which 29% were critical lures. However, when the critical lures were actually presented, the probability of recall was 95% if presented in Position 1 and 89% if presented in Position 10.

The primary purpose of the second experiment was to allow us to estimate the signal-detection parameters for each of the item types in a more traditional fashion, namely using confidence intervals. Table 2 contains two measures of sensitivity and one measure of criterion for each of the three item types from Experiment 2. Two measures of sensitivity were used for comparison:

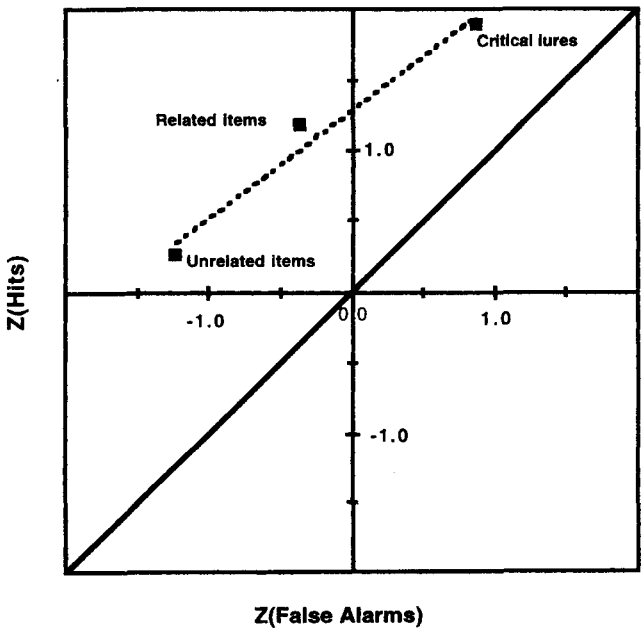

Figure 2. The receiver operator characteristic function across item types in Experiment 1.

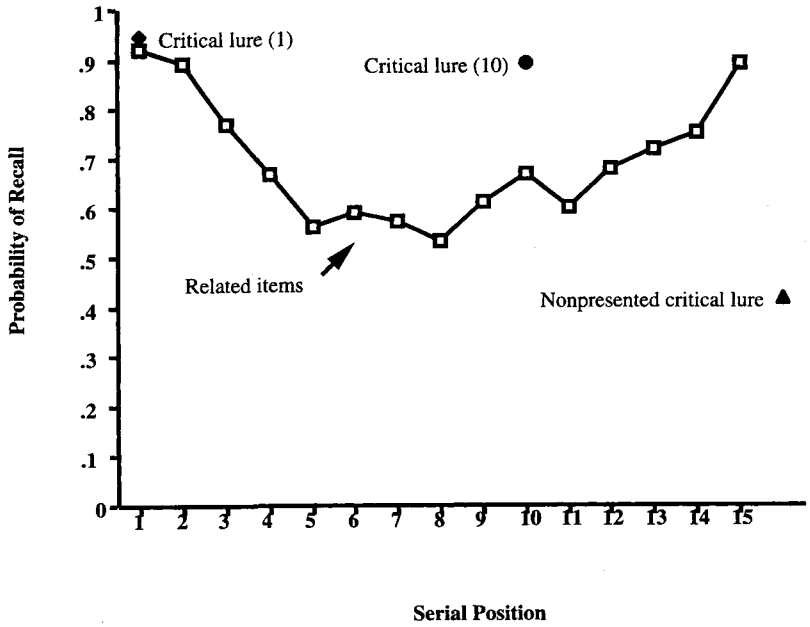

Figure 3. Probability of correct recall as a function of serial position in Experiment 2.

$d(a)$ and $d'e$. The measure $d(a)$ was used in the previous experiments because it can be calculated from a single point on the ROC curve (Macmillan & Creelman, 1991). In this case, $d(a)$ was calculated for each confidence interval and then averaged across the intervals to provide a measure for each item type. McNicol (1972) recommended using the measure of sensitivity based on the entire ROC curve and assuming unequal variance. Once the slope of a ROC curve plotted in normal–normal coordinates is determined, a value for sensitivity can be obtained when the $z(H)$ equals 0. McNicol defined $d'e$ as "twice the value of z(H) or z(FA), ignoring the signs, at the point where the ROC curve intersects the negative diagonal," because $z(H)$ and $z(FA)$ are equal when the ROC meets the negative diagonal. The measures $d(a)$ and $d'e$ produced nearly identical values, although those values were slightly lower than the corresponding ones obtained in Experiment 1. Both measures support the main conclusion of Experiment 1 that the item types are similar in sensitivity but differ substantially in criterion.

We examined the slope of the best fit regression line for each of the three item types separately. Slopes less than 1.0 indicate that the old and new distributions differ in variance. The average slope of the slopes for the three item types was 0.75, which compares favorably with the value of 0.76 from the first experiment and from the average value of 0.78 obtained by Ratcliff et al. (1992).

Discussion

We have shown that the nonpresented critical lures in the Roediger and McDermott (1995) paradigm do not behave as if they had been presented. Performance is significantly higher on every measure when the critical lures are presented compared with when they are not. Further, we have shown that the performance on critical lures is more consistent with a criterion shift than with a change in sensitivity. The notion of a criterion shift to explain the recognition data raises several important questions: (a) What mechanisms could lead to such a criterion shift, (b) how would a criterion shift explain the remember–know judgments, (c) how would a criterion shift explain the free-recall data, and (d) what do our data imply for the concept of false memories?

One mechanism for explaining the criterion shift would be to postulate that participants develop metaknowledge of the structure of the stimulus lists as the task proceeds. Casual comments from the participants made it clear that they realized that most of the lists contained highly related items. After the fact, they undoubtedly could have produced some of those themes or categories for the lists. Given that metaknowledge, one could further postulate that if an item on the recognition test were recognized as being related to one of the remembered categories, the participant would use a lower criterion for responding "old" to that item. Related items would have a higher probability of triggering one of the categories than unrelated items by construction. Further, because of the way the lists were constructed, the critical lures would have a higher probability of triggering one of the categories than the related items. Huttenlocher, Hedges, and Duncan (1991) proposed a model for estimating spatial location. In that model, participants are assumed to combine categorical and item information in a Bayesian fashion to produce estimates in ambiguous cases. The model predicts several common biases in memory and psychophysical tasks. The model would provide a mechanism for producing different criteria for the different item types in the current research.

In addition to the metaknowledge and its influence on criterion, the critical lures undoubtedly differ in a number of other respects from the related items. By the way they were chosen, they almost certainly have more high associates than the other items, and they are probably higher in frequency and a number of other characteristics as well. These structural differences could affect their

position on any underlying distribution of familiarity, both for the nonpresented items and for the presented items. Our unrelated lists were constructed from critical lures and related items from lists that were not used in the experiment. There is no pattern or theme to the items on the unrelated lists. These critical lures on the unrelated lists have the same structural properties as the critical lures from the related lists, but they are presented without the context of their related items. In the second experiment, those unused critical lures could appear as presented on the unrelated lists or could appear on the recognition tests without having been presented. Across all participants, critical lures from the unrelated lists yielded 68% hits and 27% false alarms. Related items from the same unrelated lists yielded 63% hits and 19% false alarms. The difference in false alarms is borderline significant using chi square ($p = .052$). (The chi-square test assumes independence, which is violated in this analysis but it provides a gauge of the difference.) Roediger and McDermott (1995) tested a very similar comparison in their second experiment. They tested critical lures and related items whose context lists were never presented. They obtained values of 16% for critical lures versus 11% for related items, and that difference was significant in their sample. Therefore, it appears that critical lures do have higher initial strength than other items.

A reviewer suggested an alternative model to account for our findings. The reviewer proposed that the presence of the related list context could shift both the signal and the noise distributions for critical lures and shift them more than the corresponding distributions for the related items. The signal and noise distributions for the related items also could be shifted relative to the unrelated items, but not as much as the critical lures. The differences that we observe could result from these shifted distributions rather than shifted criteria. We believe the bulk of evidence favors criteria shifts.

The first bit of evidence comes from the free-recall data. We argue later that participants may use a generate–recognize strategy during free recall. In our two experiments, the nonpresented critical lures were recalled well above chance but significantly worse than the related items at any point in the serial position curve. Recall of the nonpresented critical lures was substantially higher in the Roediger and McDermott (1995) experiments, averaging well above the recall of the related items in the middle serial positions. One possibility for the lower recall in our studies is that once participants see a presented critical lure on one of the lists, they realize how strong an item can be and they implicitly adjust their criterion for future events accordingly. Participants never see a presented critical lure in the Roediger and McDermott studies. To test this possibility, we pooled across our two experiments, and analyzed the probability of recalling a nonpresented critical lure before and after participants had experienced a presented critical lure. Participants reported the nonpresented critical lure 68% of the time when it was tested before any critical lure was presented and 32% of the time when preceded by a list containing a critical lure. The preceding analysis is confounded by list order so that the 68% is from List 1 and the 32% is from Lists 2 and 3. To examine the effect of list order alone, we looked at recall of all items in List 1 versus all items in Lists 2 and 3. The corresponding recall percentages for related items were 71%, 67%, and 68% on Lists 1, 2, and 3, respectively. Thus, we see a dramatic difference between recall performance on nonpresented critical lures before and after participants saw their first presented critical lure on a study list. That difference does not appear to be attributable to list order or any other structural effects that we can think of. The difference is consistent with an understandable change in criterion. We are not suggesting that they recognize a critical lure as a critical lure, but once they experience the very high strength of a presented critical lure, their standard or criterion changes for evaluating other items. Of course, a change in criterion would only be apparent if participants were using and shifting criteria during the task, providing evidence in favor of our interpretation.

A second and closely related argument is that Roediger and McDermott (1995) found the false-positive rate on critical lures to be equal to or greater than the hit rate on related items. In both of our experiments, the false-positive rate is lower than the hit rate. That difference could reflect the same effect described in the previous paragraph. Namely, because all of our participants eventually experience a presented critical lure, they have a different standard for judging "oldness" at recognition time than the participants in the Roediger and McDermott study, who never experienced a presented critical lure. Again, if that speculation were true, it would argue for criterion shifts rather than distributional shifts.

A third argument against the distributional shift model is the equality of d' for the three item types in both experiments. In our model relying on shifting criteria, we would expect similar values for d' across item types. With the distributional shift model, any similarity of d' across item types would be a coincidence. In some ways, an ideal false memory would yield a d' of 0. The false memory would be just as strong as if it had been presented. Any successful model of Roediger and McDermott's paradigm has to take account of the fact that a critical lure profits just as much from an actual presentation as any other item. Finally, criterion shifts would be the standard interpretation given the data that we obtained. We believe the data are more consistent with a shift in criteria rather than a shift in distributions.

The same metaknowledge, criterion differences, and distributional differences that are postulated to affect old versus new recognition could produce the observed results on the remember versus know judgments. Once the participant responded "old" to an item and was confronted with a choice between remember and know, the participant could set a new criterion for that judgment. The effects postulated previously would have similar effects on the remember–know judgments. Some investigators have argued that remember–know judgments are qualitatively different and could not be fit by a signal-detection model (e.g., Gardiner, Ramponi, & Richardson-Klavehn, 1998). However, others have argued persuasively that a two-criterion signal-detection model can account for many, if not all, of the remember–know findings given assumptions about the shape of the distributions (Donaldson, 1996; Hirshman & Henzler, 1998; Hirshman & Master, 1997). We do not mean to imply that memory is a unidimensional process, just that the remember–know judgment may involve criterion setting and that criterion could vary by item type.

To model performance on free recall requires an additional assumption. We have to postulate that during free recall, several words not on the presented list would be considered as candidates for recall. Those items would be put through something akin to a recognition process. In such a generate–recognize model, critical lures would be advantaged over other items types in two respects. Because the lists were designed by picking the 15 highest associ-

ates to the critical lure, the critical lure should be easier than any other item to generate as a candidate for recall. Because of the initial structural advantage of the critical lures as described earlier, the critical lures would also have an advantage over other items on the recognition phase. Several authors have assumed a generate–recognize model in explaining recall phenomenon, particularly allegedly improved recall under hypnosis (Dywan & Bowers, 1983; Klatzky & Erdelyi, 1985). Roediger and Payne (1985) found results that argued against such a model. They showed that manipulating criterion through instructions did not increase the number of hits in a free-recall paradigm. Erdelyi, Finks, and Feigin-Pfau (1989) replicated Roediger and Payne's results, but demonstrated that the failure to find an effect of criterion was to some extent the choice of list items that could not be guessed or generated by participants. A generate–recognize model clearly cannot work if the correct items cannot be generated. Erdelyi et al. included lists that contained items that could be guessed and obtained an effect of instructed criterion. Roediger, Srinivas, and Waddill (1989) acknowledged the results of Erdelyi et al., but argued that the effects were quite small.

To summarize the previously discussed recall studies, there does appear to be evidence for a generate–recognize model (or at least criterion effects in free recall) but only if the correct items can be generated by the participants. The effects might be small, but that smallness may reflect a low probability of generating correct items by chance rather than a weakness in the model. The Roediger and McDermott (1995) paradigm was designed in such a way that the critical lures had a very high probability of being generated by the participants by chance. The 15 presented items were chosen to be the 15 highest associates to the critical lure. If there were ever a paradigm in which a generate strategy would work (for the critical lures), this is it. Given the ease of generating the critical lures, the effect of a generate–recognize strategy should be larger than usual.

Although nonpresented critical lures were reported well above chance frequencies on free recall, we think it is critically important to remember that actually presented critical lures were recalled three to four times as often in both of the current studies. Nonpresented critical lures do not behave as if they had been presented.

As we described earlier, the intrusion of nonpresented critical lures in free recall decreases dramatically after participants experience their first presented critical lure. That decrease is consistent with a change in criterion, but a change of criterion would only affect performance if criteria played a role in recall performance.

We believe our findings provide additional clues to the understanding of false memories. What are false memories? Just about any test ever constructed to test memory yields some errors. One question to ask is whether any error on a memory test should be considered a false memory. In our view and in the view of the researchers whom we sampled, the answer is "no." Many errors are guesses, misunderstandings, confusions, and so on. How, then, can one single out which ones are false memories and which ones are not? In the current literature, there appear to be a couple of principles for arguing that a memory is false. One principle seems to be that if a condition or item type yields a high percentage of errors (false-positive results), then maybe those errors are actually false memories. How do we know it is a high percentage? One standard would be that the condition produces as many false memories as a similar condition produces true memories. Roediger and McDermott (1995) stressed the high level of false-positive responses and the similarity of false-positive responses on critical lures to the hits on related items. A thought experiment addresses this standard.

Take a standard old–new memory experiment with unrelated words. We present the list to Group A and at test time instruct them to say "old" only if they are quite sure. Group A yields 52% hits and 5% false-positive answers. Are those 5% false-positive answers false memories? In a court of law, anyone of them might wreak havoc, but a researcher would have trouble making a big fuss in the literature about the 5% errors. If that were the criterion, every memory study ever done would be about false memory. To continue, we then present the same list to Group B but instruct them to say "yes" if there is any chance at all that they might have seen the item. Group B yields 90% hits and 50% false-positive responses. That is a lot of false-positive answers. In fact, it is roughly the same percentage as Group A had for hits. Does that make Group B a good candidate for studying false memory? Most researchers probably would say "no." Group B is no more interesting than Group A; the participants' criteria were simply lowered through instructions.

A second standard would be that the participants indicate on a second judgment that they actually believe the false memories are real. The remember–know judgment or confidence ratings are candidates for that second judgment. Although a false-positive response that is labeled as having been actually remembered does seem more convincing than a false-positive response alone (and might carry more weight in a courtroom), the remember–know judgment is not tapping some infallible system of metaknowledge. Participants are rarely asked to make this judgment in ordinary circumstances, and in spite of the detailed instructions, some ambiguity must exist. As cited previously, several researchers have argued that the remember–know judgment involves the setting of another criterion and as such is participant to criterion shifts just like the primary old–new judgment. This criterion has some of the same difficulty as the previous criterion. Namely, a small percentage of remember judgments among the false positives would not interest many people, but we could carry out the same thought experiment as previously done and show that the percentage of remember judgments is no more satisfying as an indicator of "true" false memories than the percentage of false-positive memories. Confidence judgments are conceptually similar. Most signal-detection researchers assume the creation of several additional criteria to model confidence ratings (Macmillan & Creelman, 1991). These additional criteria could vary as a function of item type.

An important issue with respect to false memories is the locus of the effect. Roediger and McDermott (1995) stated that many theories have assumed the locus to be at encoding. They suggested that it could occur at encoding or retrieval. Even the prevailing vocabulary suggests an encoding locus. Papers talk about "implanting" false memories, and the concept of "recovered" memories is ubiquitous. Such terms imply an encoding locus. A criterion shift is something that happens at retrieval, but is more of a decision process than a retrieval process per se. We believe that many people would be surprised, and a bit disappointed, to learn that some forms of false memories result from criterion shifts.

So where does that leave us? Fortunately or unfortunately, the definition may well depend on the intended use. One definition may be required in a courtroom, and quite a different one may be

required for a researcher interested in doing functional magnetic resonance imaging studies on brain processes involved in false memories. For instance, a researcher who wanted to test whether false memories are stored in the same location as true memories would have a difficult time if the intended paradigm produced false memories through criterion shifts. On the basis of our data, we believe that criterion effects will play an important role in any comprehensive definition.

We are not proposing a specific model of the Roediger and McDermott (1995) paradigm. There are several models of the recognition process that include signal-detection processes such as Mandler (1980) and Yonelinas (1997). We believe that any complete model of the Roediger and McDermott paradigm will need to allow for criterion shifts.

References

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

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin, 113,* 403–439.

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

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533.

Dywan, J., & Bowers, K. (1983). The use of hypnosis to enhance recall. *Science, 222,* 184–185.

Erdelyi, M. H., Finks, J., & Feigin-Pfau, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental Psychology: General, 118,* 245–254.

Gardiner, J., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7,* 1–26.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics.* Huntington, NY: Krieger. (Original work published 1966)

Hirshman, E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9,* 61–65.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition, 25,* 345–351.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimatimating spatial location. *Psychological Review, 98,* 352–376.

Johnson, M. K., Nolde, S. F., Mather, M., Kounios, J., Schacter, D. L., & Curran, T. (1997). The similarity of brain activity associated with true and false recognition memory depends on test format. *Psychological Science, 8,* 250–257.

Klatzky, R. L., & Erdelyi, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *International Journal of Clinical & Experimental Hypnosis, 33,* 246–257.

Lindsay, D. S., & Read, J. D. (1994). Psychotherapy and memories of childhood sexual abuse: A cognitive perspective. *Applied Cognitive Psychology, 48,* 518–537.

Loftus, E. F. (1979). The malleability of human memory. *American Scientist, 67,* 312–320.

Loftus, E. F. (1993). The reality of repressed memories. *American Psychologist, 48,* 518–537.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge, England: Cambridge University Press.

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

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

McNicol, D. (1972). *A primer of signal detection theory.* Sydney: Australasian Publishing Company.

Miller, M. B., & Gazzaniga, M. S. (1998). Creating false memories for visual scenes. *Neuropsychologia, 36,* 513–520.

Murdoch, B. B. (1974). *Human memory: Theory and data.* Potomac, MD: Erlbaum.

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

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

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global models of memory using ROC curves. *Psychological Review, 99,* 518–535.

Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin and Review, 3,* 105–111.

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

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition, 13,* 1–7.

Roediger, H. L., Srinivas, K., & Waddill, P. (1989). How much does guessing influence recall? *Journal of Experimental Psychology: General, 118,* 255–257.

Schacter, D. L., Reiman, E., Curran, T., Yun, L. S., Bandy, D., McDermott, K. B., & Roediger, H. L. III. (1996). Neuroanatomical correlates of veridical and illusory recognition memory: Evidence from positron emission tomography. *Neuron, 17,* 267–274.

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.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80,* 481–488.

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

Received December 26, 1997
Revision received August 25, 1998
Accepted August 25, 1998 ■