



The role of decision criterion in the Deese–Roediger–McDermott (DRM) false recognition memory: False memory falls and rises as a function of restriction on criterion setting

Jerwen Jou, Eric E. Escamilla, Mario L. Arredondo, Liann Pena, Richard Zuniga, Martin Perez & Clarissa Garcia


To cite this article: Jerwen Jou, Eric E. Escamilla, Mario L. Arredondo, Liann Pena, Richard Zuniga, Martin Perez & Clarissa Garcia (2016): The role of decision criterion in the Deese–Roediger–McDermott (DRM) false recognition memory: False memory falls and rises as a function of restriction on criterion setting, The Quarterly Journal of Experimental Psychology, DOI: [10.1080/17470218.2016.1256416](https://doi.org/10.1080/17470218.2016.1256416)

To link to this article: <http://dx.doi.org/10.1080/17470218.2016.1256416>



Accepted author version posted online: 03 Nov 2016.
Published online: 23 Nov 2016.




[Submit your article to this journal](#) 



Article views: 6



[View related articles](#) 



[View Crossmark data](#) 

The role of decision criterion in the Deese–Roediger–McDermott (DRM) false recognition memory: False memory falls and rises as a function of restriction on criterion setting

Jerwen Jou, Eric E. Escamilla, Mario L. Arredondo, Liann Pena, Richard Zuniga, Martin Perez and Clarissa Garcia

Department of Psychology Science, University of Texas–Rio Grande Valley, Edinburg, TX, USA

ABSTRACT

How much of the Deese–Roediger–McDermott (DRM) false memory is attributable to decision criterion is so far a controversial issue. Previous studies typically used explicit warnings against accepting the critical lure to investigate this issue. The assumption is that if the false memory results from using a liberally biased criterion, it should be greatly reduced or eliminated by an explicit warning against accepting the critical lure. Results showed that warning was generally ineffective. We asked the question of whether subjects can substantially reduce false recognition without being warned when the test forces them to make a distinction between true and false memories. Using a two-alternative forced choice in which criterion plays a relatively smaller role, we showed that subjects could indeed greatly reduce the rate of false recognition. However, when the forced-choice restriction was removed from the two-item choice test, the rate of false recognition rebounded to that of the hit for studied list words, indicating the role of criterion in false recognition.

ARTICLE HISTORY

Received 12 December 2015
Accepted 28 October 2016

KEYWORDS

Criterion shift account; Deese–Roediger–McDermott false memory; Deese–Roediger–McDermott false recognition; residual experiences; two-alternative forced choice criterion-restricted; yes/no versus two-alternative choice

Since the publication of Roediger and McDermott's (1995) seminal paper, studying a type of false memory created in a lab setting has become one of the major memory research interests in nearly the past two decades. Roediger and McDermott's (1995) revived and expanded some memory experiment procedures and materials originally developed by Deese (1959) into what is now known as the Deese–Roediger–McDermott (DRM) paradigm to study false memory in a controlled lab setting. In this paradigm, subjects are presented with a list of semantically associated words (such as *bed, rest, awake, tired, dream, doze, night, peace, late, drowsy, snooze, snore, blanket, slumber, nap*) with the meanings of the words converging on a theme (*sleep*), but with the thematic word itself not presented at study. In the recall and recognition tests in their 1995 study, Roediger and McDermott found that the rate of falsely recalling or recognizing the nonstudied thematic word (henceforth *critical lure*) closely matched that of the actually

studied words (McDermott, 1996; Roediger & McDermott, 1995). This dramatic demonstration of the dismal fallibility of memory sparked an intense interest in this phenomenon in the memory research community (Bruce & Winograd, 1998; Gallo, 2010).

The robustness of the phenomenon of false memory in the DRM paradigm and its resistance to elimination have been extensively documented since 1995 (Anastasi, Rhodes, & Burns, 2000; Gallo, Roberts, & Seamon, 1997; Gallo, Roediger, & McDermott, 2001; Heit, Brockdorff, & Lamberts, 2004; McDermott, 1996; McDermott & Roediger, 1998; Neuschatz, Payne, Lampinen, & Toggia, 2001; Payne, Elie, Blackwell, & Neuschatz, 1996; Roediger, 1996; Roediger & McDermott, 2000; Toggia, Neuschatz, & Goodwin, 1999; but see also Jou, 2008; Jou & Foreman, 2007). For example, although giving strong explicit warnings and source-monitoring instructions to alert subjects to the critical lure in experiments did lower the false memory rate somewhat, the reduction was nowhere

near eliminating this effect (Gallo et al., 1997; Gallo, Roediger, et al., 2001; McDermott & Roediger, 1998; Multhaup & Conner, 2002).

The most remarkable aspect of this false memory is that subjects are not just claiming that the critical lures are familiar, but claim to have actually “remembered” the event in the sense that they can consciously re-experience all the contextual, physical, and temporal details associated with the presentation and studying of that theme word. Many studies using the DRM paradigm reported subjects’ vivid and distinctive recollections of having seen or heard “certain details” of the critical lure (Gallo, McDermott, Percer, & Roediger, 2001; Lampinen, Meier, Arnal, & Leding, 2005; Lampinen, Neuschatz, & Payne, 1999; Payne et al., 1996; Roediger & McDermott, 1995), including remembering the list positions at which the critical lures were presented and the other list words presented near them, or the gender of the speaker who delivered the critical lure (Gallo, McDermott, et al., 2001; Hicks & Marsh, 1999; Lampinen et al., 1999; Neuschatz et al., 2001; Norman & Schacter, 1997; Payne et al., 1996; Reed, 1996). Subjects also expressed a level of confidence about studying the critical lures comparable to that for the studied words (Reed, 1996; Roediger & McDermott, 1995). Because of the seemingly subjective indistinguishability of the false memory from true memory and the reality-distorting and the involuntary nature of this experience (Coren & Girgus, 1978; Hershenson, 1989), some researchers called these vivid, realistic-appearing memories of the nonexperienced events *phantom recollections* (Brainerd, Payne, Wright, & Reyna, 2003; Brainerd, Wright, Reyna, & Mojardin, 2001) or *memory illusions* (Roediger, 1996).

However, there are two different views on the cognitive processes underlying false memory in the DRM paradigm. One is a memory-based (or distribution-shift) account, and the other a decision-based (or criterion-shift) account of false memory (Miller, Guerin, & Wolford, 2011; Miller & Wolford, 1999; see also Roediger & McDermott, 1999; see Wickens & Hirshman, 2000; Wixted & Stretch, 2000, for a review). The memory-based account argues that false memory is a memory created by an implicit associative response mechanism or spreading of activation from the related concepts on the list that subjects have studied (Gallo, McDermott, et al., 2001; Gallo & Roediger, 2002; Roediger, Balota, & Watson, 2001; Roediger & McDermott, 1995; Underwood, 1965) and is the same in nature as true memory. Therefore, for example, when people claim that they saw the word “anger” (a critical

lure), the false recognition is as compelling, vivid, and impervious to strategic control (such as warning) as true memory (Gallo, McDermott, et al., 2001; Gallo & Roediger, 2002; Gallo, Roediger, et al., 2001; Roediger, 1996; Lampinen et al., 1999; Roediger & McDermott, 1999). The decision-based account, however, argues that the memory for the critical lure is derived at least in part from a criterion shift toward a more liberal position for the critical lure than for other words. Specifically, the reason why people respond “yes” to the critical lure in the recognition test is that they reply on their knowledge of the semantic structure of the lists and infer that the critical lure is probably on the list (Brainerd & Reyna, 1998; Miller, Guerin et al., 2011; Miller & Wolford, 1999). The implication of this account of false memory is that the recognition decision is not based on a compelling experience of studying the word, but on a strategic, inferential judgment. An example offered by Lampinen, Neuschatz, and Payne (1997) illustrates very well the distinction between these two accounts of false memory. A person was watching a video of an episode of someone eating at a restaurant. However, in the middle of the video presentation, the viewer dozed off for a moment and missed a small section of the whole episode, the part of paying the bill. The viewer was later asked whether she saw the diner paying the bill. She did not actually remember seeing the diner paying the bill, but she inferred that the diner did. So, she responded “yes” to the payment question just as someone who actually saw the act of paying the bill. According to Lampinen et al. (1997), the inferred “memory” does not have the compelling, vivid nature of true memory.

Based on the results of their signal detection analysis of the criteria that subjects adopted in a DRM paradigm recognition experiment, Miller and Wolford (1999) argue that the false memory is derived from a decision-criterion shift (shifting the criterion to a more liberal position on the memorial strength-of-evidence axis when judging the critical words relative to the other words), rather than from memorial evidence of the critical lure. In other words, in making a recognition decision, subjects demand less memory evidence for a “yes” decision for the critical lure than for other words. The memory-based view, on the other hand, argues that according to the logic of signal detection theory, although a criterion change in decision can result in a measured criterion shift, a measured criterion shift for the critical words does not necessarily translate into a criterion change in

the decision process. Instead, the measured criterion shift can be derived from a shift of the memory evidence distribution of the lure toward the right relative to the studied list words (for a detailed explanation, see Wixted & Stretch, 2000). Wickens and Hirshman (2000) concurred with Wixted and Stretch (2000) in that although adopting a more liberal criterion in decision results in a measured criterion shift toward a liberal bias, a measured criterion shift does not necessarily entail an actual more liberal criterion placement in the decision making. In addition, Wixted and his colleague (Stretch & Wixted, 1998; Wixted & Stretch, 2000) suggest that making within-a-trial-block, item-by-item frequent criterion shifts requires a great deal of mental effort, which subjects normally are unwilling to carry out.

Likewise, trying to refute the criterion-shift interpretation of false memory, Roediger and his colleagues (Gallo, Roediger, et al., 2001; Roediger & McDermott, 1999) argue that if the false recognition is the result of adopting a very liberal criterion, then warning the subjects on the appearance of the critical lure and against accepting it should greatly lower or eliminate the false recognition. Given that studies have shown that warnings on the nature of the critical lure produced only a moderate reduction of false memory (Gallo et al., 1997; Gallo, Roediger, et al., 2001; McDermott & Roediger, 1998), Roediger and McDermott (1999) conclude that decision criterion shift is unlikely to be the source of the false memory. However, when Miller, Guerin, et al. (2011) used what they called *criterion warning* (which expressly warned subjects that they should watch out for a word that was related to the theme of all the related words and reject that word because that word was not studied), the false-alarm rate of the critical lure was indeed substantially reduced (from .77 to .46).¹ Moreover, they found that the reduction in false alarm (FA) and the associated criterion shift was larger for the critical lure than for the related and unrelated words. Hence, they concluded that at least to a certain extent the DRM false recognition is amenable to warning and hence derived from a criterion shift.

Despite having indicated their view that false memory is unlikely to be the result of criterion shifts, neither Wixted and Stretch (2000) nor Wickens and Hirshman (2000) ruled out the possibility of a criterion shift as the source of the DRM false memory. As Wixted and Stretch (2000) stated, "Admittedly, it is within the realm of possibility that participants are willing to exert the mental effort required to adjust

the criterion item by item on the basis of category membership even though they are not inclined to . . ." (p. 376). Likewise, Wickens and Hirshman (2000) stated:

One can never completely rule out the possibility that a participant makes adjustments to decision process to accommodate local circumstances. . . . A more productive approach is to draw on the results of other studies to argue for the plausibility or implausibility of the processes underlying the different explanations. (p. 380)

Thus, the conclusion is that a measured criterion shift as presented by Miller and Wolford (1999) is not sufficient evidence to unequivocally demonstrate that false recognition is decision based, not that there is no possibility of a criterion shift being the source of false memory. Both Wixted et al. and Wickens and Hirshman (2000) have indicated that there has not been a definitive conclusion so far on the issue of a memory- versus decision-based account of false memory. Both of them agree that more research and evidence (beyond a measured bias change within the signal detection theory framework) is needed to determine whether the false recognition in the DRM paradigm is, at least to a certain degree, derived from criterion shifts.

As noted, researchers holding the memory-based view (Gallo, Roediger, et al., 2001; Roediger & McDermott, 1999) argue that resistance of the false-memory phenomenon to the influence of warnings is evidence against the idea that false memory is the result of a conscious, strategic decision-criterion adjustment. We think that this argument is based on some untested assumptions. The first assumption on which this argument is based is that people can identify the critical lure consciously. The second assumption is that criterion setting is a conscious process. We think that people may be able to identify some critical lures in the DRM paradigm when they want to minimize the FA rate as demonstrated by Neuschatz, Benott, and Payne (2003), but this process may be largely implicit or unconscious under most circumstances and therefore not amenable to conscious, strategic control (Jou & Foreman, 2007). Jou and Foreman (2007) have presented data suggesting the implicit nature of the knowledge about the critical lure and of its identification process in a training procedure using immediate corrective feedback to minimize false recall and recognition rate of the critical lures. They found that subjects could greatly lower the rate of recalling and recognizing the critical lure due to a corrective feedback but they could not explicitly identify

or produce the critical lure. If subjects cannot consciously identify the critical lures, how can they consciously raise the criterion for these probes (even though they may possibly be able to do it at an unconscious level)? Miller et al. (2011) used their criterion warning and considerably lowered the rate of FA to the critical lure. The criterion warning was in essence an instruction that explicitly and elaborately taught subjects how to identify the critical lures at test. Other studies that investigated the role of criterion setting in DRM false memory also manipulated instructions (McCabe & Smith, 2002; McDermott & Roediger, 1998; Neuschatz et al., 2001). Could such explicit warning have provided subjects knowledge they did not have before or brought the implicit knowledge to the conscious level? The warning manipulation cannot answer this question.

Ideally, the criterion-shift account can be tested without using an elaborate and explicit warning instruction. Is there a way subjects can lower the critical-lure FA rate other than being explicitly warned about the “trick” word? One cannot make a recognition decision in the yes/no (YN) test without employing a criterion of whether or not one is warned against the critical lure. A two-alternative forced-choice (2AFC) test is considered to be a criterion-free recognition test by some researchers (Egan, 1975; Green & Swets, 1966; Hicks & Marsh, 1998; Macmillan & Creelman, 1991; Zechmeister & Nyberg, 1982) although others may not agree depending on how one defines *criterion free*. In principle and under an ideal condition, a 2AFC test is criterion free—that is, if subjects choose the left item 50% of the time, and the right item 50% of the time. When subjects have a left- or right-choice bias, the 2AFC is actually not criterion free in the strict sense of the term that the bias or criterion measure is completely neutral (Jou & Flores, 2010; Jou, Flores, Cortes, & Leka, 2016; Macmillan & Creelman, 1991). That said, even though the 2AFC is not criterion free in the strict sense, there is no doubt that the role of criterion setting in a 2AFC is considerably reduced relative to that in a YN test for the reasons to be given later. Therefore, 2AFC can at least be used to manipulate the extent to which criterion plays a role in a recognition test.² This is what we did in this study. In the remaining of the article, *criterion free* is not meant in the strict sense, but in the relative sense (that is, relative to a YN test, criterion plays a lesser role in a 2AFC). The present study aimed to investigate these questions by comparing the critical-lure FA rate in a YN test

with its corresponding FA rate in a 2AFC test. We also used two novel forms of paired-items tests (described later in the introduction) to further investigate what role criterion placement plays in the DRM false recognition.

It is not uncommon to obtain different results and conclusions in memory and perception research depending on how the test is designed and conducted (e.g., Guerin, Robbins, Gilmore, & Schacter, 2012; Hicks & Marsh, 1999; Koriat, Pansky, & Goldsmith, 2011; Migo, Montaldi, Norman, Quamme, & Mayes, 2009; Naveh-Benjamin & Kilb, 2012, among others). Of particular interest to us are cases where some test procedures can tap “dormant” or implicit knowledge that can otherwise go undetected. For example, when Higham (2002) repeated the classic Thomson and Tulving’s (1970) encoding specificity experiment but used a forced recall procedure, the specific associative cue effect that Thomson and Tulving observed vanished, suggesting that subjects had memory of the original target words even under changed cue words. Similarly, McCloskey and Zaragoza (1985) found no misleading post-event information effect (in which memory of the witnessed event is supposed to be altered or impaired; Loftus, 1979; Loftus & Loftus, 1980) when they used a new distractor in a recognition test rather than the distractor presented in the misleading post-event episode, showing that memory for the originally witnessed object still existed. As well, in studying people suffering vision losses due to lesions to the occipital lobe, Weiskrantz (1980) found that these people indicated no awareness of the existence of a stimulus when a self-report test was used, but showed consistent, above-chance detection and localization of the same stimulus when a 2AFC test was used. Based on this finding, Weiskrantz suggested that there is residual vision (sometimes also known as “blind-sight”) in these people, which can be detected, but only with a 2AFC test.

Since we contrasted the YN with the 2AFC test in this study, we discuss in the following the properties of a 2AFC test that might possibly make the difference between it and the YN test in the capability for detecting “elusive” memory. In a YN test, a single test item is presented, and subjects make either an “old” or a “new” decision for the single test item. There are two factors in this decision process, the *d* prime (*d'*), or sensitivity, which measures discrimination ability, and the *criterion*, which measures a bias towards responding yes or no. The bias is affected by

motivation and is supposed to be independent of the true discriminability measure d' . One can make more Yes or No responses depending on one's considerations separate from pure evidence of prior experience with the event in the recognition test. In a 2AFC test, one must and may choose only one of the test items as "old" regardless of whether any or both items are judged to have met an absolute criterion. Thus, ideally, the criterion is supposed to be irrelevant, meaning that one does not compare the test items against an absolute strength-of-evidence criterion that one chooses to adopt. Instead, the two items are compared against each other, and the one with the higher signal value is chosen (Egan, 1975; Glanzer & Bowles, 1976; Green & Swets, 1966; Hicks & Marsh, 1998; Macmillan & Creelman, 1991; Smith & Duncan, 2004; Zechmeister & Nyberg, 1982). According to this idea, we rely on the familiarity or signal-strength *difference* (or *relative* familiarity) between the two items to make a recognition decision (Glanzer & Adams, 1990; Glanzer & Bowles, 1976; Green & Swets, 1966). Because the decision is supposed to be based on the strength *difference* between the two items rather than on an *absolute* criterion as in a YN test, the test is *assumed* to be criterion free (Egan, 1975; Green & Swets, 1966; Hicks & Marsh, 1998; Macmillan & Creelman, 1991; Zechmeister & Nyberg, 1982). Recently, however, Jou et al. (2016) have indicated that the 2AFC test is actually not criterion free in the strict sense of the word. Despite this caveat, it is fair to say that criterion setting plays a reduced role in the 2AFC test compared with its role in a YN test.

Another property of a 2AFC relevant to the present study is that the information of the two items is redundant (Jou et al., 2016). That is, in a 2AFC recognition test, if one can recognize the first item as old, then the second item has to be new. In other words, the information in one item is redundant with the information in the other item, constraining the old/new possibilities of each of the two items. McKenzie, Wixted, Noelle, and Gjurjyan (2001) expressed the same view about the old/new 2AFC. They indicate that because the two items in an old/new 2AFC are mutually exhaustive and exclusive, if one item can be identified as old, the status of the other item is determined (i.e., new). In a single-item YN recognition, on the other hand, there are four possible outcomes for two items presented as two probes: yes/yes, yes/no, no/yes, no/no. When two items are presented in a 2AFC, there can be only two outcomes: yes/no and

no/yes. This redundant relation between the two items gives an advantage to the 2AFC test over the YN test for identifying the target (Jou et al., 2016).

In Experiments 1 and 2, we compared critical-lure FA rates across YN and 2AFC tests to assess the role of criterion setting in false recognition. In this comparison, the role of criterion setting (full versus reduced criterion role across YN and 2AFC) and the number of probe items presented in a trial (one in YN versus two in 2AFC) were confounded. If the false-recognition rate of the critical lure decreases in the 2AFC compared with the YN test, is it because the criterion role is restricted in the 2AFC or because the probe pair mate in the 2AFC provides helpful clues for identifying the target item? To address this question, in Experiments 3 and 4, we contrasted the standard 2AFC with a two-alternative test in which subjects were allowed to make a free choice on the two items—that is, they could choose one of the two items, neither one, or both items. This test is in effect equivalent to presenting two YN test items simultaneously in one test trial. This manipulation achieves two goals. First, it eliminates the number-of-items confounding in the YN versus 2AFC comparison. Second, because the requirement for having to choose one and only one item in a 2AFC is removed, subjects can resume the adoption of an absolute criterion in the two-alternative free-choice test if they so desire. In this test format, they can apply an absolute criterion to each of the two test items in the two-item choice test just as they do in a YN test. This novel test will more directly and unequivocally determine whether it is the decision criterion per se or some information afforded by a pair mate that causes the critical-lure FA rate in a 2AFC to drop (if it does) from the level in a YN test. We discuss this point further in Experiment 3.

Therefore, by comparing the 2AFC test with the YN and with a free-choice version of the two-alternative test, we attempt to find out what will happen to the critical-lure FA rate when the role of criterion setting is reduced in a 2AFC test. The results will answer our question of whether subjects commit high rates of critical-lure FA in the standard YN test because they cannot distinguish the critical lures from the studied words, or because they have the knowledge (possibly implicit) to discriminate the critical lures from studied words, but choose not to use it to reduce the false-alarm rate (which would support the criterion-shift account). In short, do subjects still commit critical-lure FA at the same rate in a 2AFC as they do in the

standard YN test? Investigating the role of criterion setting by constraining what subjects can do in a test does not involve coaching them how to discriminate the critical lure from the studied words, and hence whatever response they make reflects what they originally know, uncontaminated by the “training” they receive in the warning.

There were several published studies in which 2AFC was used to investigate DRM false memory. For example, Westerberg and Marsolek (2003) compared false recognition of the critical lure across YN and 2AFC, but their purpose was different from ours. They wanted to ensure that the observed sensitivity (d') rank ordering of three types of words (d' of critical lure $< d'$ of unrelated words $< d'$ of related words) was not due to response biases associated with the judgment in the YN test. Since the 2AFC test is considered to be bias free, if they could obtain the same sensitivity of rank-ordering for these three types of words in the 2AFC test, then they could rule out the role of bias in obtaining the differential sensitivities among the three types of words. Although the FA rate of the critical lure dropped from .74 in their Experiment 1 (a YN test) to .46 in Experiment 2 (a 2AFC test), it rebounded to .70 in Experiment 3 (which also used a 2AFC test except with word frequencies of the three types of word better equated). The critical-lure FA rate was not the focus of their study, and they did not address the question of why it dropped and rebounded when tested in two 2AFC tests. We did notice an interesting pattern in their data although the authors did not discuss it: The critical-lure FA rate and its bias c measure were inversely correlated (in Experiment 1, a YN test, FA rate = .74, $c = -.70$; in Experiment 2, a 2AFC, FA rate = .46, $c = -.04$; in Experiment 3, a 2AFC, FA = .70, $c = -.68$). Thus, in that study, the critical-lure FA rate and its liberal bias went in tandem, although this aspect of the data was not a target in that study and was not commented on by the authors. Their results, however, did demonstrate the central point in Jou et al.'s (2016) study that a 2AFC test may not be always criterion free (the irony is that Westerberg & Marsolek, 2003, tried to rule out a bias interpretation of their d' rank ordering for the three types of words, and yet in the 2AFC of their Experiment 3, they actually obtained a response bias of $-.68$ for the critical words, almost the same size as they obtained in the YN test of their Experiment 1, apparently without their awareness; see their Figure 12, page 756).

Gallo and Seamon (2004) tested false recognition of the critical lure in a 2AFC test (in which the pair mate was an unrelated distractor) to find out whether subjects would falsely recognize the critical lure at above chance level when they could not perceive or recall any of the related semantic associates (in a close-to-subliminal display condition) prior to the recognition test. They found that some minimum perception or recall of the semantic associates (e.g., at least one semantic associate) was necessary for an above-chance-level false recognition of the critical lure in a 2AFC. Again, although they used the 2AFC test to study false memory, they did not compare the YN test with 2AFC, and their purpose was not to investigate the role of criterion in false memory, but rather to determine whether conscious thought of the critical lures was necessary in generating false recognition.

Weinstein, McDermott, and Chan (2010) compared the critical-lure false-recognition rate at an immediate test with that of a 7-day delayed test in a 2AFC test (in which the critical lure was paired with a studied list word). They found that the false-recognition rate for the critical lure at the immediate test was .32 as compared to a chance level at the 7-day delay test. However, the authors did not give an explanation for why they obtained a much lower critical-lure FA rate at the immediate 2AFC test than the typical FA rate that researchers normally obtain in a YN recognition test (typically above .60). Our study focuses on providing an explanation for why there is such a large discrepancy in the false-recognition rate of the critical lure across the two forms of tests on the basis of the differential roles that decision criterion plays across the two tests.

Thus, although this is not the first study examining the critical-lure FA rate in a 2AFC, to our knowledge, this is the first study to use the YN versus the 2AFC test, and the standard 2AFC versus a free-choice two-alternative test as a means of manipulating the role of decision criterion in the DRM false-recognition memory and to give a criterion-setting-based explanation for why the critical-lure FA rate is substantially lower in the 2AFC than in the YN test.

Experiment 1

In Experiment 1, we conducted a YN test closely modelling the recognition test in Roediger and McDermott's (1995) study and contrasted the results from that test with the counterpart of a 2AFC equivalent.

In the 2AFC test, a studied list word was paired with one of the distractors including the critical lure. Our goal was to change the test format but hold the learning part of these two testing conditions constant and see how much difference in the critical-lure FA rate the YN versus the 2AFC test could make.

Experiment 1 was also conducted to obtain a baseline measure of the critical-lure FA rate to compare with the corresponding measures obtained in Experiments 2, 3, and 4 when we used tests with more procedural modifications.

Method

Subjects

A total of 145 undergraduate psychology students participated in the experiment, 74 in the YN test condition, and 71 in the 2AFC test condition, for course credit.

Design and materials

The two test conditions (YN vs. 2AFC) were a between-subjects factor. The materials and procedure at the study phase were the same for the two conditions but the testing procedures were different. Twenty-four lists of DRM words (of 15 words each, not including the thematic word) were adopted from Jou, Matus, Aldridge, Rogers, and Zimmerman (2004) lists, which were in turn adopted from Roediger and McDermott's (1995) study. Odd-numbered subjects studied and were tested on half of them (i.e., 12 lists), and even-numbered subjects on the other half of the 24 lists. The procedure by which the list words were presented for study, the selection of the distractor words for the recognition test, and the procedure of presenting the test words in the YN test condition were by and large modelled after Roediger and McDermott's (1995) Experiment 1 with the following modifications. The first change was that subjects learned one list of the associated words rather than multiple lists as in the Roediger and McDermott's (1995) study before they took the recognition test. Second, at the study phase, the list words were presented one at a time in a new random order for each subject rather than in the same descending order of backward association strength of the words. This was done to eliminate the serial position effect in memory as a potential confounding variable (e.g., in a probe pair of the 2AFC, the item at the beginning of the list at presentation might receive the benefit of a primacy effect and get chosen more frequently).

The test words for a studied list were composed by following the DRM paradigm (Roediger & McDermott, 1995). They consisted of six words, of which one was the critical lure for that particular list, one word was selected from each of Positions 1, 8, and 10 of the studied list, and the remaining two words were randomly chosen from a set of 24 nonpresented words (consisting of six critical words for six nonstudied lists, and 18 other distractor words, three from each of the six nonstudied lists—one word from Position 1, one from Position 8, and one from Position 10 of each of the 6 lists). Thus, in this design, all nonpresented distractor words except the critical lure for the target list were unrelated distractors.

The same six test words were used in the 2AFC test. The pairing of the two words in the 2AFC test condition was randomly determined for each subject with the constraint that one of the two words had to be a studied word and the other a distractor word. The sides on which the target words appeared were counterbalanced across subjects. Thus, there were a total of three test pairs for each studied list of words, two of which were composed of a studied list word and an unrelated distractor, and one of which a studied list word and the critical lure of the target list. The total backward association strength of the test probes for the two test conditions should be equal because the list words and the number of list words learned in the two test conditions were exactly the same. Although the number of test items was doubled in the YN relative to the 2AFC test condition, the number of individual-word probes processed by the subjects was equivalent. Therefore, the difference was more of a physical format nature than of a cognitive one.

Procedure

Under both test conditions, the 15 words of a list were presented one at a time at a pace of 2.5 s per word in the centre of a computer monitor with a 1-s blank screen between the presentation of one word and the next. The order in which the lists were presented for learning was also random. Subjects were asked to memorize these words as best as they could. At the end of the presentation of a list of words, subjects pressed the enter key to start a 30-s backward counting task before the recognition test (in the Roediger & McDermott, 1995, study, subjects also performed an interpolated activity before the recognition test). The backward counting started with a random three-digit number, and subjects counted it down by 3 at

a time. Subjects produced the next number by typing the number onto a blank space on the screen and pressing the enter key. When the produced number was wrong, an error message was displayed accompanied by a warning tone and an instruction to re-count from the preceding number.

When the 30-s counting expired, they pressed the enter key to start the recognition test. For the YN test condition, at the recognition test, the six probe words were displayed in the centre of the monitor one at a time in a new random order for each subject. Odd-numbered subjects were instructed to press the “z” key if they judged that they saw that word at study, and the “/” key if they judged that they did not see that word. The mapping of the responses to the keys was reversed for the even-numbered subjects. They were provided with an index card with the word “Yes” on it to be put on either the left or the right side of the keyboard to mark the positive response side in case they forgot which key was for the “yes” response. There was a 1-s blank screen between the end of pressing the response key and the appearance of the next probe word. Subjects were told that accuracy of their responses was very important and that if their responses were random, the computer would detect the random pattern, and consequently they would be asked to repeat the experiment. However, nobody actually repeated the experiment. They were also told that they should not take a rest while the test word was being displayed but could take a brief break at the end of the test before they started on learning the next list of words. For the 2AFC test, the two words were displayed horizontally in the centre of the screen separated by four spaces. Subjects pressed either the “z” or the “/” key corresponding to the side of the word they judged old to indicate which word was the studied word. Other aspects of the procedure were the same as those in the YN test condition.

Results and discussion

The mean FA rate of the critical lure and both the mean hit rate of the studied list words and the FA rate of the unrelated distractors as a function of experiment and test condition are presented in Table 1. Note that the list-word hit rates were different when they were paired with the critical lure versus with an unrelated distractor.

The FA rate for the critical lure was .651 in the YN test (which was about the same as the average of .660 in Stadler, Roediger, & McDermott’s, 1999, DRM critical-

lure FA rate norm) and .295 in the 2AFC test condition (which was in the neighbourhood of the critical-lure FA rate of .32 obtained by Weinstein et al., 2010, also using a 2AFC test procedure). The percentage of drop in the critical-lure FA rate from the YN to the 2AFC was 55%. The difference was significant, $F(1, 143) = 91.87$, $MSE = .050$, $p < .0001$, $\eta^2 = .389$. The FA rate of the unrelated distractors was .024 in the YN test and .032 in the 2AFC test condition, respectively. The difference was not significant ($F < 1$). Thus, although the test method greatly affected the FA rate of the critical lure, it had no effect on the FA rate of the unrelated distractors.

The mean sensitivity measure d' for the noncritical words (with hit rate calculated from the studied list words and FA rate calculated from the unrelated distractors and the critical lure combined) was 1.94 in the YN test, and 1.95 in the 2AFC test condition.³ The difference was not significant ($F < 1$). This indicated that, unlike for the critical lure, the discrimination between the studied words and the distractor items (including unrelated distractors and the critical lure) was not affected by the test format. We then examined the noncritical words’ recognition criteria in both test conditions by measuring c .⁴ With this measure, when the criterion is neutral or unbiased, c is equal to zero. When the criterion is conservatively biased, c has a positive value, and when the criterion is liberally biased, it has a negative value (Macmillan & Creelman, 1991). The mean criterion measure c for the noncritical words was $-.225$ in the YN test, and $-.236$ in the 2AFC test condition. The difference was not significant ($F < 1$). The mean sensitivity measure d' and the mean criterion measure c of the noncritical words are presented in Table 2 (the critical lures did not have these two measures since no critical words were presented at study).⁵

To summarize, the most important finding was that the FA rate for the critical lure was cut down by 55% from the YN test to the 2AFC, indicating that restricting the use of criterion in the 2AFC accounted for a significant proportion of the drop from the critical-lure FA rate typically observed in the YN recognition test. It is remarkable that this proportion of reduction in false recognition was achieved without warning the subjects on the presence of, and against the acceptance of, the “trick” words in the test.

Experiment 2

The first purpose of Experiment 2 is to replicate the results (a large reduction of FA rate for the critical

Table 1. Mean hit and false-alarm rates as a function of experiment, test condition, and type of words.

Experiment/Test Condition	Test-Probe Type	Hit rate	False-alarm rate
<i>Experiment 1</i>			
YN test	Critical lure	—	.651 (.031)
	Studied list word	.869 (.012)	—
2AFC test	Unrelated distractor	—	.024 (.006)
	Critical lure	—	.295 (.020)
	Studied list word (paired with unrelated distractor)	.968 (.012)	—
	Studied list word (paired with critical lure)	.705 (.020)	—
	Unrelated distractor	—	.032 (.012)
<i>Experiment 2</i>			
YN test	Studied critical word	.908 (.017)	—
	Critical lure	—	.482 (.038)
2AFC test	Studied list word	.864 (.010)	—
	Related distractor	—	.137 (.014)
	Studied critical word	.964 (.012)	—
	Critical lure	—	.212 (.024)
	Studied list word (paired with related distractor)	.945 (.007)	—
	Studied list word (paired with critical lure)	.788 (.024)	—
	Related distractor	—	.082 (.009)
<i>Experiment 3</i>			
2AFC test	Studied critical word	.964 (.012)	—
	Critical lure	—	.212 (.024)
	Studied list word (paired with related distractor)	.945 (.007)	—
	Studied list word (paired with critical lure)	.788 (.024)	—
2A-FRC test	Related distractor	—	.082 (.009)
	Studied critical word	.873 (.014)	—
	Critical lure	—	.541 (.028)
RAP-FRC test	Studied list word	.810 (.01)	—
	Related distractor	—	.201 (.015)
	Studied critical word	.898 (.014)	—
	Critical lure	—	.635 (.026)
	Studied list word	.848 (.008)	—
	Related distractor	—	.190 (.013)
<i>Experiment 4</i>			
YN test	Studied critical word	.932 (.016)	—
	Critical lure	—	.785 (.038)
	Studied list word	.782 (.020)	—
	Related distractor	—	.293 (.028)
2AFC test	Unrelated distractor	—	.120 (.021)
	Studied critical word	.920 (.021)	—
	Critical lure	—	.402 (.035)
	Studied list word (paired with noncritical distractor)	.892 (.015)	—
	Studied list word (paired with critical lure)	.598 (.035)	—
	Related distractor	—	.181 (.024)
2A-FRC test	Unrelated distractor	—	.082 (.015)
	Studied critical word	.951 (.015)	—
	Critical lure	—	.854 (.023)
	Studied list word (paired with noncritical distractor)	.780 (.020)	—
	Studied list word (paired with critical lure)	.829 (.023)	—
RAP-FRC test	Related distractor	—	.435 (.035)
	Unrelated distractor	—	.121 (.019)
	Studied critical word	.931 (.016)	—
	Critical lure (paired with studied list word)	—	.886 (.034)
	Critical lure (paired with unrelated distractor)	—	.843 (.033)
	Studied list word	.826 (.047)	—
	Related distractor	—	.500 (.035)
	Unrelated distractor	—	.171 (.031)

Note: The number in the parentheses is the standard error of the mean. YN = yes/no; 2AFC = two-alternative forced-choice; 2A-FRC = two-alternative free-choice; RAP-FRC = rearranged-pairing free-choice.

Table 2. Mean sensitivity measure d' and criterion measure c as a function of experiment, test condition, and type of words.

Experiment/Test Condition	Test-Probe Type	d'	c
<i>Experiment 1</i>			
YN test	Critical lure	—	—
	Noncritical word	1.94 (.075)	-.225 (.033)
2AFC test	Critical lure	—	—
	Noncritical word	1.95 (.061)	-.236 (.026)
<i>Experiment 2</i>			
YN test	Critical word	1.19 (.103)	-.539 (.064)
	Noncritical word	2.41 (.098)	-.006 (.04)
2AFC test	Critical word	1.48 (.064)	-.289 (.036)
	Noncritical word	2.29 (.083)	-.112 (.022)
<i>Experiment 3</i>			
2AFC test	Critical word	1.48 (.064)	-.289 (.036)
	Noncritical word	2.29 (.083)	-.112 (.022)
2A-FRC test	Critical word	.640 (.060)	-.560 (.044)
	Noncritical word	1.35 (.057)	-.004 (.032)
RAP-FRC test	Critical word	.532 (.054)	-.733 (.045)
	Noncritical word	1.46 (.046)	-.059 (.029)
<i>Experiment 4</i>			
YN test	Critical word	.283 (.097)	-.914 (.080)
	Noncritical word	1.91 (.094)	.131 (.053)
2AFC	Critical word	1.02 (.088)	-.465 (.047)
	Noncritical word	1.79 (.089)	.015 (.025)
2A-FRC	Critical word	.237 (.063)	-1.11 (.047)
	Noncritical word	1.82 (.097)	.107 (.061)
RAP-FRC	Critical word	.160 (.065)	-1.09 (.049)
	Noncritical word	1.55 (.104)	-.049 (.057)

Note: The number in the parentheses is the standard error of the mean. YN = yes/no; 2AFC = two-alternative forced-choice; 2A-FRC = two-alternative free-choice; RAP-FRC = rearranged-pairing free-choice.

lure from the YN to the 2AFC test) in Experiment 1 with a test format that was modified from the DRM paradigm in more aspects than in Experiment 1. If the same result pattern is obtained, the conclusion that decision criterion contributes to false memory can be generalized to a wider range of testing conditions. Second, in the test of the DRM paradigm (and Experiment 1), the only semantically related distractor was the critical lure. All the other distractors were semantically unrelated to the theme of the target list. Some studies have indicated that criterion setting in recognition is sensitive to distractor characteristics (Benjamin, 2005; Benjamin & Bawa, 2004; Brown & Steyvers, 2005). Among other findings, these studies show that when the targets and distractors are highly discriminable (in this case, the highly dissimilar majority of distractors), a more liberal criterion tends to be adopted. When the distractors and targets are similar, a more stringent criterion tends to be employed. Thus, the similarity of the distractors to the targets seems to be one factor that may affect the FA rate of the critical lure. Indeed, Gunter, Ivanko, and Bodner (2005) found that false recognition

of critical lures decreased substantially in the related-distractor test context compared with the unrelated-distractor context. When the majority of the distractors in the DRM paradigm are unrelated, subjects may form an impression that most of the targets are related, and most of the distractors are unrelated, and, therefore, if the test probe is related, it is likely to be a studied word, leading to a high FA rate for the critical lure. On the other hand, when all the distractors are related, subjects have to rely more on the item-specific information rather than the thematic categorical information.

Another change made in Experiment 2 was to have all the 15 studied words and an equal number of related distractors tested. Some theories of recognition memory suggest that making decision in a recognition test is a process of learning to categorize probes as members of an old or a new category (Estes & Maddox, 1995; Hirshman, 1995). To correctly categorize a probe, one needs to construct an old-item distribution and a new-item distribution. Information about the old-item distribution is derived from the items on the study list, and information about the new-item distribution is derived from the distractors on the test-item list. According to Hirshman (1995), subjects need to estimate the range (the high and low points; Parducci, 1984) of the old- and the new-item distributions to be able to accurately calibrate the criterion placement. By providing the whole set of the old and new items (instead of a fragment of the distribution as in the DRM paradigm test) in Experiment 2, we think that a memory representation of the old- and new-item distributions can be formed that will more accurately reflect the actual discrepancy in familiarity between the two distributions, and consequently a reduction in the rate of the critical-lure FA from the level observed in Experiment 1 can be achieved.

Lastly, in Experiment 1, the d' and the criterion c of the critical words could not be calculated because no critical word was presented at study in the DRM paradigm. In this experiment, half of the critical words were studied, and the other half were not studied by half of the subjects, and the studied and nonstudied two halves of the critical words were reversed for the other half of the subjects. Thus, each critical word was studied by half of the subjects and not studied by the other half of the subjects. With this design, we could calculate the hit and FA rates of the critical words in the conventional way and measure sensitivity d' and criterion c for the critical

words in the YN and the 2AFC condition to see whether there was a d' and criterion change across the two test formats.

Still another question for Experiment 2 to answer is how the test format will affect the recognition of other related words compared with the critical lure. In other words, is the critical lure especially sensitive to the criterion manipulation compared with other related words? Experiment 2 will allow us to answer the question of how significant a role decision criterion plays in the recognition of the critical lure and the recognition of the other related words. To our knowledge, this experiment is the first attempt to examine how YN versus 2AFC tests possibly affect criterion placement for the critical and other related words differently. If it affects the critical words more than the other words, that will suggest that criterion plays a more important role for the former than for the latter.

Method

Subjects

A total of 117 undergraduate psychology students participated in the experiment, 57 in the YN condition, and 60 in the 2AFC condition, for course credit.

Design and materials

As in Experiment 1, there were two test conditions, the YN and the 2AFC test condition. Again, the study part of the experiment was the same as that in Experiment 1 and the same also for the two test conditions. The testing part of the experiment was changed from that in Experiment 1. Each list of the 15 semantic associates plus the thematic word (16 words) used in Experiment 1 was extended into a list of 30 semantically associated words. The twenty-four 30-word lists were adopted from Jou et al. (2004). We describe how the expansion of the list was done in the following. The additional 14 words were selected from Russell and Jenkins's (1954) associated word norms. Since Roediger and McDermott (1995) developed their word lists by taking the top 15 words in a chosen list of the Russell and Jenkins (1954) original word association norm (with some exceptions, see Roediger & McDermott, 1995), we simply chose our additional 14 words from each corresponding list in Russell and Jenkins's norm by generally picking the next 14 words from each list (with some exceptions, e.g., to skip a word or two in a list to avoid already

used words). Thus, the 30 words in each extended list (one of the 30 words was the critical word) were ordered in descending semantic association value with the exception of the critical word. The critical word was positioned as the first word for the odd-numbered lists, but as the second word for the even-numbered lists. For half of the subjects, the 15 odd-numbered words on each list were presented for study, and the 15 even-numbered words were not presented, and were used as distractors at test. Therefore, for this half of the subjects, the critical words in the odd-numbered lists (which were the first words on the lists) were presented, but the critical words in the even-numbered lists (which were the second words) were not presented. For the other half of the subjects, the even-numbered words on each list were presented, and the odd-numbered words were not presented. Hence, half of the subjects studied the critical words on the odd-numbered lists, and the other half of the subjects studied the critical words on the even-numbered lists. With this method of selecting targets and distractors, when the critical word was presented, the top noncritical word on the list (with the highest backward association strength) was omitted from presentation and was used as a distractor. When the critical word was not presented and used as a critical lure, the top noncritical word on the list was presented instead and used as a target.

For the YN test, there were 30 test words, 15 of which were studied, and 15 not studied. For half of the lists, one of the 15 studied words was the studied critical word. For the other half of the lists, one of the 15 nonstudied words was the critical lure. For the 2AFC, there were 15 test pairs. A test pair in the 2AFC was composed of a studied word randomly selected from the 15 studied words, and a nonstudied distractor randomly selected from the 15 nonstudied words of the list. For a list of which the critical word was studied, the studied critical word was randomly paired with a nonstudied word. For a list of which the critical word was not studied, the critical lure was randomly paired with a studied word. The target and distractor were assigned to the left and right sides of a pair with equal probability, and target and distractor sides were counterbalanced across subjects. The 15 words studied and the 15 words not studied and the 30 test words used in the YN and the 2AFC test conditions were identical, and hence the backward association strength was kept equal across the two test conditions.

Procedure

The study and testing procedures were the same as those in Experiment 1 except that there were a total of 30 test trials for the YN test and 15 test trials for the 2AFC for each list of words. The order of presenting the lists for study and the test trials for each list of words in both test conditions was random.

Results and discussion

The mean hit rate and the FA rate as a function of experiment, test format, and type of probe words are presented in Table 1. The most notable change in recognition performance from the YN test to the 2AFC test was the large drop in the FA rate of the critical lure from .482 in the YN test to .212 in the 2AFC test. The reduction was 56% and significant, $F(1, 115) = 37.16$, $MSE = .057$, $p < .0001$, $\eta^2 = .244$. In contrast, the hit rate for the critical words showed a significant increase from .908 in the YN test to .964 in the 2AFC test, $F(1, 115) = 7.00$, $MSE = .013$, $p = .009$, $\eta^2 = .058$. The drop in the FA rate accompanied by an increase in hit rate for the critical word indicated an improvement in the discriminability between the studied and nonstudied critical words from the YN to the 2AFC test. The mean d' and c for the critical and the noncritical words (composed of the related studied words and distractors) are presented in Table 2. Consistent with the decrease in FA rate and the increase in hit rate of the critical word, its mean d' increased from 1.19 in the YN test condition to 1.48 in the 2AFC test condition (after a downward adjustment by a factor of $1/\sqrt{2}$), and the increase was significant, $F(1, 115) = 5.86$, $MSE = .424$, $p = .017$, $\eta^2 = .048$.

The criterion c measure for the critical words was $-.539$ (liberally biased) in the YN, and $-.289$ (comparatively less liberally biased) in the 2AFC test condition. The difference was significant, $F(1, 115) = 11.79$, $MSE = .154$, $p = .0008$, $\eta^2 = .093$. This indicated that a more liberal criterion placement in the YN test than in the 2AFC was a contributing factor in the higher FA rate observed for the critical lures in the YN test. When the criterion was shifted to a less liberal point due to a change in the test format, the FA rate of the critical lure was cut down by more than half. The fact that the mean c was not zero in the 2AFC indicates that 2AFC is not criterion free (Jou et al., 2016). However, the significant reduction in the liberal bias from $c = -.539$ to $-.289$ did indicate that criterion indeed played a

lesser role in the 2AFC than in the YN test. More important, the co-occurrence of reduction in liberal bias and the reduction of the critical-lure false recognition from the YN to the 2AFC indicates that liberal criterion placement is a contributing factor in the generation of the false-recognition memory.

The mean hit rate of the studied list words increased significantly from .864 in the YN to .945 in the 2AFC test condition, $F(1, 115) = 27.45$, $MSE = .005$, $p < .0001$, $\eta^2 = .193$. However, the mean hit rate of .945 was calculated from pairs that did not involve the critical lure as a pair mate. The hit rate of the studied list words when paired with the critical lure was .788, which meant that about 20% of time, subjects chose the critical lure over the studied list word (which was the FA rate of the critical lure in this test condition). The FA rate for the related distractors was .137 in the YN test, and .082 in the 2AFC test condition. The reduction in the FA rate was 40% and significant, $F(1, 115) = 10.67$, $MSE = .008$, $p = .001$, $\eta^2 = .085$. However, the mean d' for the noncritical words actually dropped from 2.41 in the YN to 2.29 in the 2AFC condition although this decrease was not significant ($F < 1$). Thus, it seemed that 2AFC was especially helpful for improving the discrimination between studied and nonstudied critical words but had no significant effect on the discrimination of the noncritical semantic associates.

The criterion c for the noncritical semantic associates was $-.006$ (almost unbiased) in the YN, and $-.112$ (moderately liberally biased) in the 2AFC test condition. The difference was significant, $F(1, 115) = 5.55$, $MSE = .059$, $p = .020$, $\eta^2 = .046$. Thus, the change of the test format from YN to 2AFC affected the criterion placement for the critical words (from $-.539$ to $-.289$, decreasing in the liberal bias) in the opposite direction to that of the noncritical words (from $-.006$ to $-.112$, increasing in the liberal bias) and affected the former to a greater extent than the latter. As noted, for the critical words, even though subjects' response in 2AFC was not completely criterion free, they adopted a less liberal criterion in the 2AFC than in the YN counterpart, and this raising of criterion was accompanied by a large drop in the FA rate of the critical lure. Also important to note was that the recognition of the critical words and the corresponding c measure were affected by the test-induced criterion shift in the predicted manner, but those of the noncritical words were not. In sum, although both the YN and the 2AFC tests in Experiment 2 were considerably modified from the format in

Experiment 1, the most important result—that is, a substantial drop in the critical-lure FA rate from the YN to the 2AFC test—was replicated.

Finally, as expected, the critical-lure FA rate in the YN condition of Experiment 2 (.482) was significantly lower than the counterpart in Experiment 1 (.651), $F(1, 129) = 12.11$, $MSE = .075$, $p = .0007$, $\eta^2 = .086$, and, similarly, the FA rate of the critical lure in the 2AFC of Experiment 2 (.212) was significantly lower than the counterpart in Experiment 1 (.288), $F(1, 129) = 6.08$, $MSE = .030$, $p = .015$, $\eta^2 = .045$.

Experiment 3

The exact cause of the remarkable drop in the rate of false recognition from the YN to the 2AFC test needs to be further investigated. More specifically, was the drop in FA rate of the critical lure due to the forced-choice mechanism that restricted the adoption of an absolute criterion, or due to the paired studied word providing the test takers additional, qualitatively different information, which made the discrimination between the true and false memory more accurate and lowered the critical-lure FA rate (Cho & Neely, 2013; Kroll, Yonelinas, Dobbins, & Frederick, 2002; Smith & Duncan, 2004). Brainer, Reyna, Wright, and Mojardin (2003) suggest that when a word highly related to a test lure is recalled, the memory of that word can lead to the rejection of the lure, a phenomenon known as recall-to-reject. Could the reason for the substantial drop in the FA rate for the critical lure in the 2AFC be that the related target word in the pair provided information or some types of clue otherwise unavailable to the subjects, which led to the more frequent rejection of the critical lure? In other words, when subjects saw the studied word juxtaposed to the critical lure, did the evidential contrast between a studied and a nonstudied word make subjects realize that they did not study the critical lure? Unfortunately, this question cannot be answered in a standard 2AFC test. In a standard 2AFC, presenting two test probes is coupled with the forced-choice requirement, thus confounding the additional information factor (if any) with the forced-choice factor (which restricts the use of an absolute criterion). If the drop in the FA rate of the critical lure in the 2AFC is caused by restricting the use of an absolute criterion, and not by any additional evidence the pairing target word may provide, then allowing the use of criteria (by allowing a free choice) in a two-word probe test should bring the FA rate of the critical

lure back to its YN test level. However, if the paired target word plays a crucial role in bringing down the critical-lure FA rate in the 2AFC due to its providing additional discriminating evidence, then lifting the choice restriction (i.e., by allowing subjects to set their criteria) in a two-item recognition test will not make the critical-lure FA rate rebound to its YN test level. To separate these two factors, we designed two test formats in which two items were presented as probes but the forced-choice requirement (whereby subjects must and may choose only one item) was removed.

This experiment also aimed to answer another important question. The rate of acceptance in recognition for the critical lure was shown to be lower than true recognition rate both in this study and in some other studies (e.g., Jou, 2011; Jou & Flores, 2013; Jou et al., 2004; Miller et al., 2011; Miller & Wolford, 1999), indicating that true memory is stronger than false memory, at least in the contexts of these studies. Since in a 2AFC, the choice is based on the relative-strength criterion, not an absolute-strength criterion, subjects will understandably choose the studied word more frequently than the critical lure. What difference in the critical-lure FA rate will it make if subjects are allowed to use the absolute criterion in a two-item probe test? If the criterion use restriction is lifted, and the FA to the critical lure is increased to the level equal to the hit rate of the studied word, then that increase in the critical-lure FA rate will provide a convincing measure of the role that criterion setting plays in false recognition of the critical lure.

In this new test design, we used the 2AFC results in Experiment 2 as the control condition results (these experiments were conducted at the same time) and constructed two other two-probe recognition tests as comparison conditions. The structure of the first of the other two test formats was the same as that of the 2AFC in Experiment 2. The only difference was in the instruction. The new instruction told subjects that they could make any of four possible responses: choosing both, choosing neither, choosing the left item, and choosing the right item. Essentially, this converts the 2AFC into two single-probe YN tests. As in a case of two single-word YN tests, there are four possible responses: yes–yes (or both old), yes–no (left item old), no–yes (right item old), and no–no (both new). This would allow the use of absolute criteria in accepting and rejecting the probes just as in the case of a YN test. The only difference of this test procedure from

the YN test is that two words are simultaneously presented, and one word is old, and the other is new. A comparison in the level of the critical-lure FA rate between this test condition and the standard 2AFC in Experiment 2 will allow us to determine the role of decision criterion in the observed false recognition while holding constant the factor of presenting a studied word side by side with the critical lure. We termed this test procedure two-alternative free choice. Unlike the two-alternative free-choice condition, the second of the two new test procedures used a test design in which one third of pairs was composed of two old items, one third was composed of two new items, and one third was composed of one old and one new item. This way of pairing the two probes made the test more equivalent to the simultaneous presentation of two single-probe YN tests—that is, either of the two probes can be either old or new, generating four possible combinations: yes–yes, yes–no, no–yes, no–no. We termed this test condition *rearranged-pairing free choice*. A comparison of the critical-lure FA rates across the control and these two new forms of two-probe test should provide an unconfounded measure of the role of criterion in the false recognition of the critical lure in a two-probe test format.

Method

Subjects

One hundred and seven introductory psychology students participated in the two-alternative free-choice condition, and 112 in the rearranged-pairing free-choice condition for partial fulfilment of the course requirement.

Design and materials

The control condition for this experiment was the 2AFC condition used in Experiment 2. The study materials in Experiment 3 were exactly the same as those in Experiment 2. Also, as in the 2AFC of Experiment 2, there were 15 test pairs for each learned list of words. The 15 test pairs in the two-alternative free-choice condition were constructed in exactly the same way as in Experiment 2 (the only difference was a change in the instruction). In the rearranged-pairing free-choice condition, one third of the 15 test pairs (5 pairs) were composed of both-new items, one third (5 pairs) of both-old items, and one third (5 pairs) of one new and one old item. In both test conditions, in the case of one old item paired

with one new item, the sides on which the old and the new item were presented were counterbalanced across subjects. In the rearranged-pairing free-choice condition, in the case of both-old and both-new pairs, the two words were assigned to the two sides randomly. The probability with which the critical word was assigned to the three types of rearranged pairs (both-old, both-new, one-old/one-new) was equal. A studied critical word could be paired either with a studied list word to make a both-old item, or with a nonstudied distractor to make a one-old/one-new item. A critical lure could be paired with either a studied list word to make a one-old/one-new item or with a nonstudied distractor to make a both-new item. Thus, in the rearranged-pairing free-choice condition, the individual test words were the same as those in the two-alternative free-choice (and 2AFC) condition. The only difference was that the same words were re-paired.

In calculating hit rates, if both old items in the both-old pairs were accepted, the response counted for two hits. In calculating FA rates, if both new items in the both-new pairs were falsely accepted, the response counted for two FAs. For the one-old/one-new pairs, a “both-old” response counted for one hit and one FA, and a correct choice of the old item counted for one hit, and a false choice of the new item counted for one FA. The important difference between the two-alternative free-choice and rearranged-pairing free-choice test conditions was that there was a redundant relationship in item old/new status between the two test probes (Jou et al., 2016) in the two-alternative free-choice test (i.e., if the left item was old, the right item must be new, and vice versa) whereas there was not such a relationship between the two test items in the rearranged-pairing free-choice condition (i.e., the old/new status of one item is independent of the old/new status of the other item).

Procedure

The procedure in the study part of the experiment was exactly the same as that in Experiment 2. The instructions for the tests were different. In both the two-alternative free-choice and rearranged-pairing free-choice conditions, subjects were told that:

If you remember that you studied both words, you press the “b” key; if you remember that you studied neither word, they press the “n” key; if you remember that you studied left-side word, they press the “z” key; and if you remember that you studied the right-side word, you press the “/” key.

In neither the two-alternative free-choice nor the rearranged-pairing free-choice were subjects told how the pairs were made up. They were given a sheet with the layout of the keyboard displayed and the four response keys conspicuously marked to make sure that they familiarized themselves with the locations of the four response keys. The 15 test pairs for a list of words were presented one at a time in a completely random order. Other aspects of the procedure were the same as in the other two experiments.

Analysis

In data analysis where multiple pairwise comparisons between means were made, in this experiment as well as in Experiment 4, both Newman-Keuls and Tukey post hoc tests were carried out. It turned out that the result patterns from these two post hoc tests were the same in those multiple pairwise comparison cases. Therefore, we reported only the test results from the Newman-Keuls test.

Results and discussion

The mean hit and FA rates of the critical words and of the noncritical words (studied list words and nonstudied distractors) obtained under the two-alternative free-choice and rearranged-pairing free-choice, along with those of the control condition (the 2AFC condition in Experiment 2), are presented in Table 1. The mean d' and c as a function of experiment and type of words are presented in Table 2. The most important results were the rebounds of the critical-lure FA rate from .212 in the 2AFC to .541 (a 155% increase) in the two-alternative free-choice, and to .635 (a 200% increase) in the rearranged-pairing free-choice condition (which was close to that of .651 in the YN test of Experiment 1). An analysis of variance (ANOVA) showed that the overall difference in FA rate was significant, $F(2, 276) = 51.52$, $MSE = .07$, $p < .0001$, $\eta^2 = .272$. A post hoc Newman-Keuls test showed that the three means were significantly different from each other. These results clearly indicated that the primary cause of the large drop of the critical-lure FA rate in the 2AFC was the restriction imposed on setting an absolute criterion in making a decision about the critical lure rather than the studied pair mate providing helpful discriminating information. In other words, pairing a studied word with the critical lure by itself was not sufficient to make subjects reject the critical lure. It was the

“being forced to choose one and only one word” rule that made them reject the critical lure. The critical-lure FA rate of .635 in the rearranged-pairing free choice was significantly higher than that of .541 in the two-alternative free choice probably due to the pairing of one target and one distractor in the latter (the same as in the 2AFC), which might have provided some subtle clues to the subjects that when one item was old the other was likely to be a new item, and vice versa. On the other hand, in the rearranged-pairing free-choice condition, the old/new status of one item was independent of that of the other item (the right item could be either old or new regardless of the left item’s old/new status), making the two-item recognition test more closely resemble two single-item YN probes presented simultaneously. The pairs that contained a critical lure were broken down into the pair type where the critical lure was paired with a studied word and the pair type where the critical lure was paired with a nonstudied word. The mean critical-lure FA rate of the former type (.633) was not significantly different from the mean critical-lure FA rate of the latter type (.618; $F < 1$). Again, the significance of the finding is that subjects were willing to accept the critical lure at a high rate when they were “allowed” to do so, even when they could “see” that the critical lure was weaker than the studied word. This is very strong evidence of the role of criterion setting.

The mean d' of the critical words for the 2AFC, two-alternative free choice, and rearranged-pairing free choice was 1.48, 0.64, and 0.53, respectively. The overall difference was significant, $F(2, 276) = 58.37$, $MSE = .329$, $p < .0001$, $\eta^2 = .297$. A Newman-Keuls test showed that the mean d' of 1.48 of the 2AFC condition was significantly higher than the other two mean d' s with the latter two means not significantly different from each other. However, the numerical difference between .64 (two-alternative free choice) and .53 (rearranged-pairing free choice) was in the direction consistent with the rearranged-pairing free choice showing a significantly higher rate of FA to the critical lures than the two-alternative free choice.

The mean criterion measure c for the critical words was $-.289$, $-.561$, and $-.733$, for the 2AFC, two-alternative free-choice, and rearranged-pairing free-choice conditions, respectively. The overall difference was significant, $F(2, 276) = 20.53$, $MSE = .189$, $p < .0001$, $\eta^2 = .129$. A Newman-Keuls test showed that all three c s were significantly different from one another. The covariation of the degree of liberalness

of the bias as measured by the three *c*s with the level of critical lure's FA is consistent with the notion that decision criterion is a contributing factor in false recognition of the critical lure. That is, the most liberal criterion produced the highest critical-lure FA rate, the second most liberal criterion produced the second highest critical-lure FA rate, and the least liberal criterion of the three produced the lowest critical-lure FA rate.

The FA rate of the noncritical distractors was .082 for the 2AFC, .201 for the two-alternative free-choice, and .190 for the rearranged-pairing free-choice conditions, respectively. The overall difference was significant, $F(2, 276) = 17.00$, $MSE = .018$, $p < .0001$, $\eta^2 = .110$. A Newman-Keuls test showed that the FA rate in the 2AFC was significantly lower than the FA rates in the two-alternative free-choice and the rearranged-pairing free-choice conditions, with the latter two means not significantly different from each other. Thus, although the overall FA rates for the noncritical distractors were much lower than that of the critical lures, criterion setting seemed to increase the FA rate of the noncritical distractors in a manner similar to that of the critical lures.

The mean d' for the noncritical words was 2.29, 1.35, and 1.46 for the 2AFC, two-alternative free-choice, and rearranged-pairing free-choice conditions, respectively. The overall difference was significant, $F(2, 276) = 59.01$, $MSE = .318$, $p < .0001$, $\eta^2 = .300$. A Newman-Keuls test showed that the mean d' of 2.29 of the 2AFC condition was significantly higher than the other two mean d' s, with the other two mean d' s not significantly different from each other. The mean criterion measure c for the noncritical words was $-.112$, $-.004$, and $-.059$ for the 2AFC, two-alternative free-choice, and rearranged-pairing free-choice conditions, respectively. The overall difference was marginally significant, $F(2, 276) = 2.68$, $MSE = .087$, $p = .071$, $\eta^2 = .019$. The pattern of the three mean *c*s indicated that subjects were actually slightly liberal in the 2AFC condition, but generally neutral or unbiased in the other two conditions. Thus, the criterion pattern across the three test conditions for the noncritical words suggested that unlike for the critical words, there was not much bias in the decision for the recognition of the noncritical words, especially under test conditions where subjects could set an absolute criterion, and that the format of testing did not seem to affect the bias in the recognition of the noncritical words in the same manner and to the same extent as it did that of the critical words.

Experiment 4

One departure from the standard DRM paradigm in the methods of Experiments 1, 2, and 3 that might have put the false memory generated in this study at an advantage compared with false memories created in other studies was that subjects were tested for their memory after they studied only one list of words. In the standard paradigm, they study multiple lists of words before being tested (e.g., in Experiment 1 of Roediger & McDermott, 1995, subjects studied six lists of words before the recognition test). This could conceivably have made the discrimination between the studied words and the critical lures easier. Moreover, in Experiments 2 and 3, the distractors used were all semantically related words, which, as the results already showed, could lower the critical-lure FA rate. In this experiment, two important changes were made to eliminate these factors that might possibly have weakened false memory relative to that in the standard DRM paradigm. The changes were, first, that the study and test materials were made the same as those used in the standard DRM paradigm (and as in Experiment 1) except that half of the critical lures were presented; second, subjects learned six lists of words (as in Roediger & McDermott's, 1995, Experiment 1) before being tested.

Again, unlike what we did in Experiment 1 (presented none of the critical words), in this experiment, we presented half of the critical words at study. This was necessary for computing the criterion c (as well as d') for the critical words. Also, we did not present the list words in the fixed order of descending backward association strength because this presentation order necessarily would have presented the critical word at the first position, making it exceptionally prominent in memory due to the primacy effect. Moreover, this presentation order would have introduced a confounding factor, the serial position effect in memory. Instead, we presented the 15 words (for half of the lists, one of them was the critical word) in a random order in order not to create a privileged memory status for the critical word and not to introduce a serial position effect for other words to confound the test results.

There were four conditions, all using the same learning and testing materials but different testing formats: (a) a YN recognition condition closely following the DRM paradigm; (b) a 2AFC condition; (c) a two-alternative free-choice condition (similar to the second condition in Experiment 3); and (d) a rearranged-pairing

free-choice condition (similar to the third condition in Experiment 3). Thus, in this experiment, we modified some procedures in Experiments 2 and 3 that might have possibly enhanced the discriminability of the critical lures from the studied words to find out whether we would obtain essentially the same result pattern as that in the earlier experiments (i.e., the fall and rise of the critical-lure FA rate depended on restricting or not restricting criterion setting), even though we expected that the overall false-recognition level of the critical lures would be raised in this experiment from the previous levels.

Method

Subjects

A total of 172 introductory and upper level psychology course students at University of Texas–Rio Grande Valley (UTRGV) participated in this experiment: 41 in the YN, 46 in the 2AFC, 41 in the two-alternative free-choice, and 44 in the rearranged-pairing free-choice conditions to fulfil a partial requirement or to receive some extra credit for the course.

Materials, design, and procedure

There were four test conditions in this experiment: a YN, a 2AFC condition, a two-alternative free-choice condition, and a rearranged-pairing free-choice condition. The test condition was a between-subjects factor. The learning part was the same for all four conditions. For the YN test condition, the materials, design, and procedure, for the most part, were the same as those in Experiment 1. An important change was that, instead of the critical words never presented as in Experiment 1, in this experiment, for half of the lists, the critical words were presented, and for the other half of the lists, they were not presented, for the reason noted earlier. The 12 lists of DRM words were divided into odd- and even-numbered lists. For half of the subjects, the critical words of the odd-numbered lists were presented, but those of the even-numbered lists were not presented. For the other half of the subjects, the critical-word-presented and not-presented lists were reversed. When the critical word was presented, a randomly selected word from the three list words (words at Positions 1, 8, and 10 on the list, which are normally chosen as the target list words at test in the DRM paradigm) was omitted from presentation and used as a related distractor at test in lieu of the critical lure. So, for a critical-word-presented list, the three target words in the six-item

probe set were the presented critical word and two studied list words (randomly selected from the 1st, 8th, and 10th words on the studied list). The three distractors were the one remaining word among the 1st, 8th, and 10th words (after two of them were chosen as studied words) and two unrelated distractors from nonstudied lists. For the lists in which the critical word was not presented, three of the six test probes were studied list words (Words 1, 8, and 10 on the list). The three distractors were the critical lure of the studied list and two list words from nonstudied lists (and hence unrelated to the studied theme). These two unrelated distractors were randomly chosen from a set of 24 nonpresented words (consisting of six critical words from six nonstudied lists, and 18 other words, three from each of the six nonstudied lists, one word from Position 1, one from Position 8, and one from Position 10 of each of the six lists) by following the DRM distractor selection protocol.

The whole set of 12 lists was partitioned into two halves of six lists each by randomly selecting six lists anew for each subject, and with these six lists constituting one round of learning and testing. The remaining six lists constituted the second round of learning and testing. The order in which the six lists in a learning/testing round were presented for studying was random for each subject. There were 36 test words totally (six from each list) in one round of testing. Of the 36 test words, 15 were presented list words, three were presented critical words, three were critical lures (associated with three studied lists), 12 were unrelated distractors (from nonstudied lists), and three were nonstudied related distractors (in lieu of the critical lures) from the studied lists with the critical words presented.

Each word was presented for 1.5 s for studying with a 1-s interpresentation interval. The 15 words were presented in a random order for the reason noted earlier. After studying a list, subjects performed backward counting for 25 s (interpolated activities were performed in the DRM procedure), and then studied another list, with another 25 s counting, and so on for six lists.

In the 2AFC condition, each of the three presented words (either three list words from Positions 1, 8, and 10, or the two randomly selected words from this triplet plus the presented critical word) was randomly paired with a distractor. The three distractors were either two unrelated words plus the critical lure (if the critical word was not presented), or two unrelated words plus one of the three words in the triplet (1st, 8th, and 10th words on the list) randomly selected

to be omitted from presentation if the critical word was presented. So, the related distractor of the three distractors could be either the critical lure (if the critical word was not presented) or one of the three words in the triplet of the related words (if the critical word was presented).

The testing procedure for the 2AFC test was the same as that in the 2AFC condition of Experiment 2 (subjects must and might choose only one word). For the two-alternative free-choice condition, the only difference from the 2AFC condition was that subjects were given complete discretion in choosing or rejecting any word or words of the test pair, as in the same condition of Experiment 3. In the rearranged-pairing free-choice condition, the individual test words were the same as those in the two-alternative free-choice condition except that the pairing was rearranged so that one third of the pairs were both-new, one third both-old, and one third one-new/one-old pairs. The instruction in this condition was identical to that in the two-alternative free-choice condition.

Results and discussion

The mean hit and FA rates of different types of probes obtained under each condition are presented in Table 1. The mean d' and c of different types of words under each condition are presented in Table 2. Of most interest are the critical-lure false-recognition rates across the four test conditions. The critical-lure mean FA rate was .785 in the YN, .402 in the 2AFC, .854 in the two-alternative free-choice, and .860 in the rearranged-pairing free-choice conditions. An ANOVA indicated that the overall difference was significant, $F(3, 168) = 52.55$, $MSE = .041$, $p < .0001$, $\eta^2 = .484$. However, a Newman-Keuls test showed only a significant difference between the mean of the 2AFC condition (.402) and the rest of the means, with the latter three means not significantly different from each other. One notable result was that the mean FA rate of the critical lure in the YN condition (.785) numerically exceeded the hit rate of the studied word (.782), although the difference was not significant ($F < 1$). Also, to assess the relative strength of the critical lure and the studied word in the 2AFC, we compared the FA rate of the critical lure (.402) with the hit rate of the studied word (.598). The difference was significant, $F(1, 45) = 8.00$, $MSE = .110$, $p = .007$, $\eta^2 = .151$. This indicated that when the two words were pitted against

each other for subjects' acceptance, the studied word won out over the critical lure.

Under the two-alternative free-choice condition, when the same items in those pairs were judged, but not with one pitted against the other in this case (since both items in a pair could be accepted), remarkably, the mean FA rate of the critical lure (.854) was numerically higher than the mean hit rate of the studied list word (.829), even though the difference was not significant ($F < 1$). So, when the mutually exclusive relationship between the two items in a pair was removed (which is to say, when choosing one and only one item restriction was lifted), the critical-lure FA rate matched, if it did not exceed, the studied-word hit rate. The same analysis was done for the two words in the rearranged-pairing free-choice condition. Again, the mean FA rate of the critical word (.886) was numerically higher than the mean hit rate of the studied list word (.826), although not significantly higher, $F(1, 42) = 1.58$, $MSE = .049$, $p = .215$, $\eta^2 = .013$.

At any rate, the basic pattern of results of Experiment 3 was replicated, although in this case, the overall FA rate of the critical lure was considerably increased from the level in Experiment 3 as expected, and the difference between the FA rate of .854 in the two-alternative free-choice and .860 in the rearranged-pairing free-choice conditions was not significant. By having subjects learn multiple lists, the discriminability of the critical lures from the studied words was lowered, but the effect of criterion setting was not diminished. The only difference between the 2AFC and two-alternative free-choice was that in the former condition, subjects were restricted from using their own criterion (the absolute criterion), whereas in the latter they were allowed to set their own criterion. This change from restricting criterion setting to not restricting it raised the critical-lure FA rate by more than 100% in this experiment. Once again, Experiment 4 replicated the crucial result of Experiment 3—that is, pairing a studied word with a critical lure did not lead to rejecting the critical lure. It was the restriction on criterion setting (subjects being forced to choose one and only one item) in a 2AFC that made all the difference in the false-recognition rate of the critical lure.

The mean d' for the critical words was .281 in the YN, 1.02 in the 2AFC, .237 in the two-alternative free-choice, and .160 in the rearranged-pairing free-choice conditions. The overall difference was significant, $F(3, 168) = 26.12$, $MSE = .275$, $p < .0001$, $\eta^2 = .318$. A Newman-Keuls test showed that only

the mean d' of the 2AFC condition (1.02) was significantly higher than the other three means, with the latter three means not significantly different among themselves, which was consistent with the 2AFC condition producing the lowest rate of critical-lure FA. The critical-lure mean criterion measure c was $-.914$ in the YN, $-.465$ in the 2AFC, -1.11 in the two-alternative free-choice, and -1.09 in the rearranged-pairing free-choice conditions. An ANOVA indicated that the overall difference was significant, $F(3, 168) = 29.32$, $MSE = .138$, $p < .0001$, $\eta^2 = .344$. A Newman-Keuls test showed that the mean c of $-.465$ of the 2AFC condition was significantly less liberal than $-.914$ of the YN condition, which was in turn significantly less liberal than -1.09 of the rearranged-pairing free-choice, and -1.11 of the two-alternative free-choice conditions with the last two means not significantly different from each other. Again, the critical-lure mean c measures were consistent with its FA rates in that the least liberal criterion (in the 2AFC condition) was associated with the lowest FA rate, and the most liberal criteria (in the two-alternative free-choice and rearranged-pairing free-choice conditions) were associated with the highest critical-lure FA rates. To reiterate, although the 2AFC test was not criterion free, it did produce the lowest liberal response bias and the lowest critical-lure FA rate among the four tests.

The FA rates of the related distractors were compared across the four test conditions to see whether there was a similar pattern to that of the critical lures. The mean FA rate was .293 for the YN, .181 for the 2AFC, .435 for the two-alternative free-choice, and .500 for the rearranged-pairing free-choice pairs. The overall difference was significant, $F(3, 168) = 22.67$, $MSE = .04$, $p < .0001$, $\eta^2 = .288$. A Newman-Keuls test showed that the rearranged-pairing free-choice mean (.500) and that of the two-alternative free choice (.435) were not significantly different from each other, but both significantly higher than the YN mean (.293), which was in turn significantly higher than the 2AFC mean (.181). Thus, it seemed that the related distractor was affected by criterion setting in the similar way to the critical lure. On the other hand, the effect of test condition on the FA to the unrelated distractor was noticeably attenuated compared with that of the related distractors, although the overall difference was significant (mean of YN = .120; of 2AFC = .082; of two-alternative free choice = .121; of rearranged-pairing free choice = .171), $F(3, 168) =$

2.74, $MSE = .021$, $p = .045$, $\eta^2 = .047$. A Newman-Keuls test showed that the mean FA rate of the unrelated distractors in the 2AFC (.082) was significantly lower than that of the rearranged-pairing free choice (.171), with that of the two-alternative free-choice (.121) and that of the YN (.120) conditions not significantly different from either of the former two means or from each other.

We computed and compared d' s for the noncritical words across the four conditions, using the related and unrelated distractors combined as new items (distractors) for calculating the FA rate. The mean d' was 1.92 for the YN, 1.79 for the 2AFC, 1.82 for the two-alternative free-choice, and 1.55 for the rearranged-pairing free-choice conditions. The overall difference was marginally significant, $F(3, 168) = 2.47$, $MSE = .425$, $p = .064$, $\eta^2 = .042$. The Newman-Keuls test indicated that this marginally significant difference derived from the difference between the highest mean d' of 1.92 (YN) and the lowest one of 1.55 (rearranged-pairing free-choice). Thus, as far as the noncritical words were concerned, the 2AFC test condition did not provide a clear discrimination advantage relative to the other test conditions. One interesting finding was that the mean recognition performance for the noncritical words was numerically worse by a substantial margin in the rearranged-pairing free choice ($d' = 1.55$) than in the YN condition ($d' = 1.92$). A comparison across the four test conditions on the criterion measure c for the noncritical words (mean of YN = .131; of 2AFC = .015; of two-alternative free choice = .107; of rearranged-pairing free choice = $-.049$) indicated that the overall difference was significant, $F(3, 168) = 2.75$, $MSE = .108$, $p = .045$, $\eta^2 = .046$. Although no strong bias toward either a liberal or a conservative direction was obvious across the four tests, the Newman-Keuls test indicated that the difference between the mean c of $-.049$ of the rearranged-pairing free choice, and that of .131 of the YN condition just hit the threshold point of significance.

In sum, even when we followed the DRM paradigm very closely, we replicated the crucial pattern of results that we obtained in the first three experiments. When multiple lists were learned before the recognition test (and hence it greatly increased the memory load), and when two thirds of the distractors was rendered unrelated, the overall performance in discrimination between true and false memories declined, but the effect of criterion setting was undiminished.

General discussion

Decision criterion is an inseparable part of memory measures (Koriat & Goldsmith, 1996; Koriat et al., 2011; Hirshman, 1995; Koriat et al., 2011; Miller & Dobbins, 2014). In recall, allowing for the use of judgmental discretion on the quality of output from memory tends to reduce the quantity of recall, but increase the accuracy of recall (Koriat et al., 2011). In other words, when one is given the freedom of setting an output criterion, it typically improves recall accuracy. Conversely, when people are forced to output a certain minimal amount of information from memory, accuracy decreases (Koriat et al., 2011). We showed a reversed pattern of the effect of criterion use in the recognition of the critical lure. When subjects were given the discretion for accepting or rejecting the test probes, their discrimination performance for the critical lure markedly dropped. When they were restricted from using their own criterion, their discrimination performance improved.

The dramatic demonstration of false memory in the DRM paradigm has attracted much attention from many memory researchers since 1995. One controversial issue that has still not been settled is whether this false memory results from a criterion shift in judging the critical lures relative to judging other words. A mere criterion measure in the signal detection framework failed to resolve the issue due to the ambiguous meaning that can be ascribed to the criterion measure. An alternative way to test the criterion-shift account is using warning instructions. The rationale is that if a warning has no effect on the false-recognition rate of the critical lure, the false recognition is not due to a criterion shift. In our opinion, using warning to test the criterion-shift account is still not the ideal way to resolve this issue decisively. Regardless of whether the explicit warning instruction is effective, there can be alternative interpretations. If it is not effective, one can argue that criterion shift (or a corrective criterion adjustment) may take place at an unconscious, but not at a conscious, level (Jou & Foreman, 2007). If it is effective, one may argue that the effect is a demand characteristic (subjects are taught to do something they cannot do under normal circumstances). We think that one of the most important contributions of this study is that we manipulated the use of criterion to assess its effect on critical-lure FA rate without explicitly telling subjects to watch out for the critical lure and to reject it. We devised test conditions to assess the role of

criterion setting in false recognition of the critical lure in which subjects themselves initiated the corrective adjustment.

To our knowledge, this study is the first to use a 2AFC as a means to manipulate the role of criterion and provide an explanation for the large critical-lure FA reduction under this test. Previous studies using 2AFC (e.g., Weinstein et al., 2010) did not use it for the purpose of manipulating the role of criterion and did not provide an explanation for why the critical-lure FA rate was lower in the 2AFC test. In addition, we designed a two-alternative free-choice test where the forced-choice feature of the 2AFC was turned off (which switched the criterion setting mechanism back on) and showed that when subjects were allowed to use an absolute criterion, they had a strong tendency to claim the critical lure as a studied word even when the actually studied word was presented side by side with the critical lure, and even when they could distinguish between the two words when they were forced to. To our knowledge, this study is the first to manipulate criterion setting by switching between a forced-choice and a free-choice mode in a two-alternative test.

This manipulation also addressed a related question concerning the nature of the difference between the YN and the 2AFC as instruments for measuring recognition memory (Cho & Neely, 2013; Kroll et al., 2002; McKenzie et al., 2001; Smith & Duncan, 2004; Yonelinas, Hockley, & Murdock, 1992). The question is whether providing a pair mate in a 2AFC provides any useful discriminating information for the target. We designed two 2-item free-choice tests in Experiments 3 and 4 to find out whether the increased discriminability of the critical lure in the 2AFC was due to the restriction on using criterion or due to the pair mate providing helpful discriminating information. We found that presenting a studied word juxtaposed with the critical lure in a test pair in itself did not seem to provide much of a useful clue for the rejecting of the critical lure. In addition, the rearranged-pairing free-choice condition in Experiment 3 revealed that when the old/new relationship between the left and right items was rendered independent of each other just as two probes in a YN test, the critical-lure FA rate further increased, and so did the liberalness of the criterion. This indicated that the large reduction in the critical-lure FA rate in the 2AFC compared with that in the YN test condition observed in Experiments 2 and 4 and in other studies cited earlier can be confidently attributed to the

restriction of the role of criterion in the 2AFC. This demonstration presented a stronger case for the role of criterion setting in DRM false recognition than a warning instruction. Ironically, however, the same results can also be construed to demonstrate how strong a lure the critical word was for the subjects. Whenever they were allowed to, they would accept the lure word as studied, suggesting their inability to raise their “subjective criterion” in the YN test to the point of rejecting the majority of critical lures.

Another important difference in our way of assessing the role of criteria than in some other studies (e.g., Miller et al., 2011; Miller & Wolford, 1999) is that we did not compare the critical word’s criterion with the noncritical word’s criterion, but compared the critical word’s criterion with its own criterion under two different test conditions in which the learning condition was held constant. This avoids the problem of the alternative interpretation that the measured criterion shift can possibly have resulted from differential memory distribution shifts for the two types of words.

In Experiment 4, we addressed the concern that the false memory that we generated in the first three experiments might not be the same as that in the standard DRM paradigm because we did not have subjects learn multiple lists before the recognition test, and we did not use unrelated distractors in Experiments 2 and 3. The results of Experiment 4 showed that multiple-list learning and the use of unrelated distractors elevated the false-recognition rate to the point of numerically exceeding the studied-word hit rate but did not change the effect of criterion setting. In all four experiments, the FA rate of the critical lures rose and fell in accordance with the availability and unavailability of the mechanism for criterion setting. Moreover, we demonstrated that true and false memories responded to the test format change (criterion-use restriction manipulation) in noticeably different ways, with false memory being more sensitive to this manipulation than true memory.

In our experiments, subjects might not be aware of how they were or were not affected by the criterion, and yet we showed that their FA rate for the critical lure changed depending on whether or not they used their own criterion. One critical question was: Do they have the knowledge to distinguish the critical lure when they are not told in a warning instruction such as a criterion warning on how to make the discrimination (Miller et al., 2011)? One possible reason why subjects commit high rates of false alarms to the critical lure may be that they do not have the

knowledge needed to distinguish the critical lure from studied words. Another possible reason is that they have the knowledge (possibly implicit) but do not apply it under certain circumstances. The fact that the critical-lure FA rate dropped by a remarkably large proportion from the YN to the 2AFC test suggested that they did have that knowledge, but that they chose not to use it, or were unable to use it in the YN test and in the two-alternative free-choice test conditions.

The present finding is consistent with results from a number of studies showing that the critical-lure FA response time is longer than that of the hit response to the studied words (Cabeza, Rao, Wagner, Mayer, & Schacter, 2001; Jou, 2011; Jou et al., 2004), which suggests that false memory is discriminable from true memory at some level. Jou (2011) and Jou and Foreman (2007) hypothesize that the nature of the knowledge for distinguishing between the critical lures and the studied words as reflected in the response time may be implicit and therefore can be manifested only in response latency but not in a conscious choice made in the YN test. The present study, however, further showed that subjects could use this knowledge to make a change at the choice level under the 2AFC test condition. When the 2AFC is imposed, subjects have to rely on a relative, instead of an absolute criterion, to make the choice. This forces them to make use of whatever sensory, memory information they may possess for discriminating the studied from the nonstudied word. As shown by Weiskrantz (1980) in the case of the “blind sights” by using a 2AFC test, there are residual visual experiences. The present study shows a form of “residual” memory similar to the residual visual experience in that it reveals itself only in a 2AFC test.

So, can we conclude on whether the false-recognition phenomenon is memory (associative activation) based, or criterion based? Or are there both an associative activation component and a criterion-setting component in producing the false recognition? If both components contribute into the critical-lure false recognition, can one component be distinguished from the other? The finding that the critical lure’s false-recognition rate was lower than the studied list-word’s hit rate when they were paired together in the 2AFC condition (e.g., .212 vs. .788 in Experiment 2; .402 vs. .598 in Experiment 4) was consistent with the response time findings from YN tests in the literature in which the FA to the critical lure was slower than the hit to the studied list word

(Cabeza et al., 2001; Jou, 2011; Jou et al., 2004). It is also consistent with Jou's (2008) finding of the critical lure's recall latency being longer than that of the studied words. All these findings seem to point to one underlying source of the difference between the critical lure and the studied word—that is, the critical lure has a lower memory activation level than the studied list word. The sensitive response time measure detects that difference, and so does a 2AFC test. In that sense, the response time and 2AFC measures can be considered more “objective” measures. We suggest that the positive recognition response rates (whether correct or incorrect) in a 2AFC test reflect the memory strengths of the test probes relative to other probes. In our view, the part of the critical-lure FA rate that is above the FA levels of other distractors in the 2AFC derives from its associative activation. In other words, the fact that the critical lure emerges as a stronger item than other distractors in a relatively criterion-free test is due to its stronger associative activation than those of other distractors. On the other hand, the lower critical-lure FA rate in a 2AFC relative to the hit rate of the studied word reflects its lower memory activation level in comparison to the studied word. The jump in the critical-lure FA rate from the level in a 2AFC to one on a par with that of a studied word in the YN, or in the two-alternative free-choice, or the rearranged-pairing free-choice test condition is brought about by a criterion setting in the subjective judgment. In other words, the discrepancy in the critical-lure FA levels between the 2AFC and the other tests is due to the latter memory measures being “contaminated” by the influence of a subjective or metacognitive judgment. So, the answer to the question of whether DRM false recognition is memory or criterion based is that the critical lure as a super-distractor is memory based, but that the elevation of the critical lure from a super-distractor to the status of a studied word is probably criterion based.

A related question to answer before concluding the discussion is if false memory is at least in part based on criterion setting, how can this conclusion be reconciled with reports of the vivid, compelling, and detailed experiences of studying the critical lure that appear no different from that of true memory? The answer lies in how the experience is measured. It has long been known that self-report often fails to truthfully reflect the underlying mental processes or the real causes of behaviours, and that the self-reported contents often differ from more objectively

observed data (Nisbett & Wilson, 1977). As mentioned above, false memory can be measured more subjectively (by using certain types of recognition and recall tests or phenomenological reports) or more objectively (by using response time measure or 2AFC test). It is not too surprising that the findings from these different ways of measuring it are not exactly the same.

Notes

1. A detailed comparison between the warning against accepting the critical lure used in Miller et al. (2011) and in other studies (e.g., Anastasi et al., 2000; McDermott & Roediger, 1998; Neuschatz et al., 2001) did not reveal fundamental differences in nature or language between these warnings. If there was any difference, Miller et al.'s gave more explicit and detailed guidance on how to identify the critical lure than others. The effects of warning found in other studies were smaller than those in Miller et al. In Anastasi et al. (2000), the false recognition rate of the critical lure dropped from .51 in the no-warning to .45 in the warning condition. In Neuschatz et al. (2001), the critical-lure false recognition rate dropped from .66 in the no-warning to .58 in both the moderate and the strong warning conditions.
2. The criterion c in a 2AFC is calculated from the hit rate and false-alarm rate of either the left-side items only or the right-side items only in the same way as it is calculated in a yes/no test. Only when 50% of the responses is a left choice, and 50% a right choice, is the criterion c measure equal to zero (Macmillan & Creelman, 1991). Note that just as in a yes/no recognition test, the hit and false-alarm rates calculated this way are based on separate sets of items (i.e., hit rate based on the half set with targets on the left-hand side, and the false-alarm rate based on the other half set with distractors on the left-hand side) and therefore the two rates are not redundant.
3. To avoid an infinite z value in computing the d 's, all hit and false-alarm rates were corrected by adding 0.5 to the frequency of hits or false alarms, and dividing this adjusted frequency by $N + 1$ where N was the number of old or new trials (Snodgrass & Corwin, 1988). The d' for a 2AFC was computed the same way as for a YN test (taking the z score difference between the proportion of hits and that of false alarms) except for a downward adjustment of that value by a factor of $1/\sqrt{2}$. Thus, d' for a 2AFC = $1/\sqrt{2} (z_{\text{Hit}} - z_{\text{FA}})$. For more information about computing the d' for a 2AFC test, see Macmillan and Creelman (1991, pp. 120–121). The false-alarm rate used to calculate the d' for the list words was defined as the proportion of distractors that was erroneously recognized as studied words in the YN test and the proportion of pairs in the 2AFC for which the distractor pair mate was chosen.
4. The value for c was computed by $c = -.50 (z_{\text{Hit}} + z_{\text{FA}})$ (Macmillan & Creelman, 1991).

5. Both d' and c are normally calculated from two response rates, the hit and the FA rates. Likewise, to measure d' and c of the critical words, the hit rate of the critical words must be available, in addition to its FA rate. When the critical words are never presented, its hit rate cannot be calculated. Some researchers, however (e.g., Arndt and Hirshman, 1998), used the critical lure's FA rate as its "hit" rate and the FA rate of the unrelated distractors as its FA rate to calculate the critical lure's d' . But other researchers (e.g., Miller et al.; 2011; Miller & Wolford, 1999; and Westerberg & Marsolek, 2003, among others) defined and computed the d' and c of the critical words in the normal way (presenting the critical words for half of the lists to obtain the hit rate, and not presenting the other half to obtain the FA rate). As Westerberg and Marsolek (2003) stated in Footnote 2: "It is important to clarify how the measures in Arndt and Hirshman (1998) differ from the more typical measures of sensitivity in recognition memory. For both human performance and simulations from MINERVA2 (Hintzman, 1988), Arndt and Hirshman calculated "sensitivity" for critical words that was more a measure of "false-recognition sensitivity" than a measure of true-recognition sensitivity. In their measure, hit rates were the probabilities of "old" responses to previously unrepresented critical words, that had associated related-word lists that were presented during encoding, and false-alarm rates were the probabilities of "old" responses to previously unrepresented critical words that had associated related-word lists that were not presented during encoding. (None of the critical words were actually presented during encoding.) This reflects a tendency to falsely remember an unrepresented critical word because its list had been presented previously. This is an interesting measure, but it is different from sensitivity as typically conceptualized and measured in recognition memory research. It does not reflect the ability to discriminate whether critical words had been presented previously, which is crucial to this discussion" (p. 750).

Acknowledgements

We are grateful to two anonymous reviewers for their very careful reviews and helpful comments on earlier drafts of this article. We thank Andy Torres for help with data collection for Experiment 4.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Anastasi, J. S., Rhodes, M. G., & Burns, M. G. (2000). Distinguishing between memory illusions and actual memories using phenomenological measurements and explicit warnings. *The American Journal of Psychology*, 113, 1–26. doi:10.2307/1423458
- Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanation from a global matching perspective. *Journal of Memory and Language*, 39, 371–391. doi:10.1006/jmla.1998.2581
- Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about apparently nonparametric measures. *Memory & Cognition*, 33, 261–269. doi:10.3758/BF03195315
- Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, 51, 159–172. doi:10.1016/j.jml.2004.04.001
- Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language*, 48, 445–467. doi:10.1016/S0749-596X(02)00501-6
- Brainerd, C. J., & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, 71, 81–129. doi:10.1006/jecp.1998.2464
- Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review*, 110, 762–784.
- 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. doi:10.1037/0278-7393.27.2.307
- Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 587–599. doi:10.1037/0278-7393.31.4.587
- Bruce, D., & Winograd, E. (1998). Remembering deese's 1959 articles: The zeitgeist, the sociology of science, and false memories. *Psychonomic Bulletin & Review*, 5, 615–624. doi:10.3758/BF03208838
- Cabeza, R., Rao, S. M., Wagner, A. D., Mayer, A. R., & Schacter, D. L. (2001). Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proceedings of the National Academy of Sciences*, 98, 4805–4810.
- Cho, K. W., & Neely, J. H. (2013). Null category-length and target-lure relatedness effects in episodic recognition: A constraint on item-noise interference models. *The Quarterly Journal of Experimental Psychology*, 66, 1331–1355. doi:10.1080/17470218.2012.739185
- Coren, S., & Girgus, J. S. (1978). *Seeing is deceiving: The psychology of visual illusions*. Hillsdale, NJ: Erlbaum.
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, 58, 17–22. doi:10.1037/h0046671
- Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York, NY: Academic Press.
- Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1075–1095. doi:10.1037/0278-7393.21.5.1075
- Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, 38, 833–848. doi:10.3758/MC.38.7.833
- Gallo, D. A., McDermott, K. B., Percer, J. M., Roediger, H. L. (2001). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 339–353. doi:10.1037/0278-7393.27.2.339

- Gallo, D. A., Roberts, M. A., & 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. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language*, 47, 469–497. doi:10.1016/S0749-596X(02)00013-X
- Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review*, 8, 579–586. doi:10.3758/BF03196194
- Gallo, D. A., & Seamon, J. G. (2004). Are nonconscious processes sufficient to produce false memories? *Consciousness and Cognition*, 13, 158–168.
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 5–16. doi:10.1037/0278-7393.16.1.5
- Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 21–31. doi:10.1037/0278-7393.2.1.21
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Guerin, S. A., Robbins, C. A., Gilmore, A. W., & Schacter, D. L. (2012). Retrieval failure contributes to gist-based false recognition. *Journal of Memory and Language*, 66, 68–78. doi:10.1016/j.jml.2011.07.002
- Gunter, R. W., Ivanko, S. L., & Bodner, G. E. (2005). Can test list context manipulations improve recognition accuracy in the DRM paradigm? *Memory*, 13, 862–873.
- Heit, E., Brockdorff, N., & Lamberts, K. (2004). Strategic processes in false recognition memory. *Psychonomic Bulletin & Review*, 11, 380–386. doi:10.3758/BF03196586
- Hershenson, M. (1989). *The moon illusion*. Hillsdale, NJ: Erlbaum.
- Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1105–1120. doi:10.1037/0278-7393.24.5.1105
- Hicks, J. L., & Marsh, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, 6, 117–122. doi:10.3758/BF03210818
- Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, 30, 67–80. doi:10.3758/BF03195266
- Hintzman, D. L. (1988). Judgment of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528–551. doi:10.1037/0033-295X.95.4.528
- Hirshman, E. (1995). Decision processes in recognition memory: Criterion shift and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 302–313. doi:10.1037/0278-7393.21.2.302
- Jou, J. (2008). Recall latencies, confidence, and output positions of true and false memories: Implications for recall and meta-memory theories. *Journal of Memory and Language*, 58, 1049–1064. doi:10.1016/j.jml.2007.12.003
- Jou, J. (2011). Conscious and unconscious discriminations between true and false memories. *Consciousness and Cognition*, 20, 828–839. doi:10.1016/j.concog.2010.10.022
- Jou, J., & Flores, S. (2010). *Is the two-alternative forced recognition test really criterion-free?* Paper presented at the 51st annual meeting of the psychonomic society in St. Louis, MO.
- Jou, J., & Flores, S. (2013). How are false memories distinguishable from true memories in the Deese-Roediger-McDermott paradigm? A review of the findings. *Psychological Research*, 77, 671–686. doi:10.1007/s00426-012-0472-6
- Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in two-alternative-forced-choice recognition: Is the test criterion-free? *Acta Psychologica*, 167, 30–44.
- Jou, J., & Foreman, J. (2007). Transfer of learning in avoiding false memory: The roles of warning, immediate feedback, and incentive. *The Quarterly Journal of Experimental Psychology*, 60, 877–896. doi:10.1080/17470210600831184
- Jou, J., Matus, Y. E., Aldridge, J. W., Rogers, D. M., & Zimmerman, R. (2004). How similar is false recognition to veridical recognition objectively and subjectively? *Memory & Cognition*, 32, 824–840.
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in strategic regulation of memory accuracy. *Psychological Review*, 103, 490–517.
- Koriat, A., Pansky, A., & Goldsmith, M. (2011). On output-bound perspective on false memories: The case of the Deese-Roediger-McDermott (DRM) paradigm. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 297–326). New York, NY: Psychology Press.
- Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General*, 131, 241–254. doi:10.1037/0096-3445.131.2.241
- Lampinen, J. M., Meier, C. R., Arnal, J. D., & Leding, J. K. (2005). Compelling untruth: Content borrowing and vivid false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 954–963. doi:10.1037/0278-7393.31.5.954
- Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1997). Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology*, 16, 181–224.
- Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1999). Source attributions and false memories: A test of the demand characteristics account. *Psychonomic Bulletin & Review*, 6, 130–135. doi:10.3758/BF03210820
- Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge: Harvard University Press.
- Loftus, E. F., & Loftus, G. R. (1980). On the permanence of stored information in the human brain. *American Psychologist*, 35, 409–420.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.
- McCabe, D. P., & Smith, A. D. (2002). The effect of warnings on false memories in young and older adults. *Memory & Cognition*, 30, 1065–1077. doi:10.3758/BF03194324
- McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, 114, 1–16. doi:10.1037/0096-3445.114.1.1

- McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, 35, 212–230. doi:10.1006/jmla.1996.0012
- 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. doi:10.1006/jmla.1998.2582
- McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes-no and forced-choice tasks. *Journal of Experimental Psychology: General*, 130, 140–155. doi:10.1037/0096-3445.130.1.140
- Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, 62, 365–386.
- Miller, M. B., & Dobbins, I. G. (2014). Memory as decision making. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences*, V. (pp. 551–563). Cambridge, MA: MIT Press.
- Miller, M. B., Guerin, S. A., & Wolford, G. L. (2011). The strategic nature of false recognition in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1228–1235. doi:10.1037/a0024539
- Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, 106, 398–405. doi:10.1037/0033-295X.106.2.398
- Multhaup, K. S., & Conner, C. A. (2002). The effects of considering nonlist sources on the deese-roediger-McDermott memory illusion. *Journal of Memory and Language*, 47, 214–228. doi:10.1016/S0749-596X(02)00007-4
- Naveh-Benjamin, M., & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: The case of remember/know judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 194–203. doi:10.1037/a0025256
- Neuschatz, J. S., Benott, G. E., & Payne, D. G. (2003). Effective warnings in the deese/roediger and McDermott false memory paradigm: The role of identifiability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 35–41. doi:10.1037/0278-7393.29.1.35
- 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. doi:10.1080/09658210042000076
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259. doi:10.1037/0033-295X.84.3.231
- Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, 25, 838–848. doi:10.3758/BF03211328
- Parducci, A. (1984). Perceptual and judgmental relativity. In V. Sarris & A. Parducci (Eds.), *Perspectives in psychological experimentation* (pp. 135–149). Hillsdale, NJ: Erlbaum.
- 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. doi:10.1006/jmla.1996.0015
- Reed, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review*, 3, 105–111. doi:10.3758/BF03210749
- Roediger, H. L. (1996). Memory illusions. *Journal of Memory and Language*, 35, 76–100. doi:10.1006/jmla.1996.0005
- Roediger, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and the arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association.
- 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. doi:10.1037/0278-7393.21.4.803
- Roediger, H. L., & McDermott, K. B. (1999). False alarms and false memories. *Psychological Review*, 106, 406–410. doi:10.1037/0033-295X.106.2.406
- Roediger, H. L., & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science*, 9, 123–127. doi:10.1111/1467-8721.00075
- Russell, W. A., & Jenkins, J. J. (1954). *The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff word association test*. (Tech Rep. No. 11, Contract N8 ONR 66216, Office of Naval Research). Minneapolis, MN: University of Minnesota.
- Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 615–625. doi:10.1037/0278-7393.30.3.615
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117, 34–50. doi:10.1037/0096-3445.117.1.34
- Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, 27, 494–500. doi:10.3758/BF03211543
- Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1379–1396.
- Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, 86, 255–262. doi:10.1037/h0029997
- Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, 7, 233–256. doi:10.1080/741944069
- Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, 70, 122–129. doi:10.1037/h0022014
- Weinstein, Y., McDermott, K. B., & Chan, J. C. K. (2010). True and false memories in the DRM paradigm on a forced choice test. *Memory*, 18, 375–384. doi:10.1080/09658211003685533
- Weiskrantz, L. (1980). Varieties of residual experiences. *Quarterly Journal of Experimental Psychology*, 32, 365–386. doi:10.1080/14640748008401832
- Westerberg, C. E., & Marsolek, C. J. (2003). Sensitivity reductions in false recognition: A measure of false memories with stronger theoretical implications. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 747–759. doi:10.1037/0278-7393.29.5.747
- Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comment on miller and wolford (1999) and roediger and McDermott (1999). *Psychological Review*, 107, 377–383. doi:10.1037/0033-295X.107.2.377

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, 107, 368–376. doi:10.1037/0033-295X.107.2.368

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 18, 345–355. doi:10.1037/0278-7393.18.2.345

Zechmeister, E. B., & Nyberg, S. E. (1982). *Human memory: An introduction to research and theory*. Monterey, CA: Brooks/Cole.