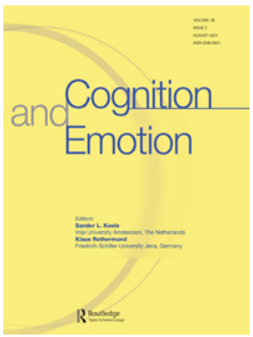

Cognition and Emotion

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/pcem20

Investigating a bias account of emotional false memories using a criterion warning and force choice restrictions at retrieval

Lauren M. Cooper, Datin Shah, Imane Moucharik & Zainab Munshi

To cite this article: Lauren M. Cooper, Datin Shah, Imane Moucharik & Zainab Munshi (28 Jul 2024): Investigating a bias account of emotional false memories using a criterion warning and force choice restrictions at retrieval, Cognition and Emotion, DOI: 10.1080/02699931.2024.2379824

To link to this article: https://doi.org/10.1080/02699931.2024.2379824

COGNITION AND EMOTION
https://doi.org/10.1080/02699931.2024.2379824

Routledge
Taylor & Francis Group

Investigating a bias account of emotional false memories using a criterion warning and force choice restrictions at retrieval

Lauren M. Cooper, Datin Shah, Imane Moucharik and Zainab Munshi

Department of Psychology, City, University of London, London, UK

ABSTRACT

Here, we add to the debate as to whether false recognition of emotional stimuli is more memory-based or more bias-based. Emotional false memory findings using the DRM paradigm have been marked by higher false alarms to negatively arousing compared to neutral critical lure items. Explanation for these findings has mainly focused on false memory-based accounts. However, here we address the question of whether a response bias for emotional stimuli can, at least in part, explain this phenomenon. In Experiment 1, we used a criterion warning, previously shown to increase more conservative responding and reduce false recognition. Experiment 2, we employed a two-alternative-forced choice test, which minimises the role of criterion setting. In both experiments, we compared false alarms to negative and neutral critical lures. We observed a significant decrease in false recognition rates for both negative and neutral critical lures under the conditions of forced choice restriction and criterion warning. However, despite these conditions, negative items, compared to their neutral counterparts, still consistently provoked a higher degree of false recognition. The discussion that follows presents an exploration of both memory-based accounts and criterion-setting explanations for the enhanced emotional false memory finding.

ARTICLE HISTORY
Received 16 January 2024
Revised 13 March 2024
Accepted 9 July 2024

KEYWORDS
Emotion; false memory; DRM paradigm; response biases

Understanding the interaction between emotion and memory has been of significant interest in the field of cognitive psychology. Such interactions are often examined in laboratory research using a study/test procedure whereby the emotional content of studied materials is varied. This variance typically occurs on two distinct dimensions; valence, an emotional value ranging from positive to negative, and arousal; the intensity of the material, ranging from low to high. Several findings have shown that memory is most enhanced when materials are negative and highly arousing (see Grider & Malmberg, 2008) compared to neutral. The explanation for this seemingly better performance for emotionally salient materials has been the subject of much debate. On the one hand, enhanced emotional memory has been attributed to activation in the amygdala (Labar & Cabeza, 2006), to the ability to capture attentional resources (Cahill et al., 1995; Cahill & McGaugh, 1998; Talmi et al., 2007a; Vuilleumier, 2005), the distinctiveness of emotional materials (Talmi et al., 2007b), and the ability to bind emotional stimuli to context (Mather & Nesmith, 2008). Whilst research has considered the complex relationship between these stimuli and neural-based explanations (Schümann et al., 2018; Talmi, 2013), they all predict a memory-based account for the enhanced emotional memory effect.

Researchers have also posited that the superior performance for emotional stimuli may reflect

CONTACT Lauren M. Cooper Lauren.Cooper@city.ac.uk Department of Psychology, City, University of London, London, EC1V 0HB, UK
Supplemental data for this article can be accessed online at https://doi.org/10.1080/02699931.2024.2379824.

response biases favouring these items (Dougal & Rotello, 2007; Grider & Malmberg, 2008; Thapar & Rouder, 2009). This could be a result of the propensity to guess more when unsure about the emotional item, or a lowered criterion setting when considering the evidence needed to classify an emotional item as old. For example, Thapar and Rouder analysed sensitivity and response bias for negative, positive, and neutral words in a memory recognition study testing older and younger adults. Negative and positive words were lower and higher in valence respectively but both were higher in arousal compared to neutral words. For younger adults, they found a more liberal response bias for emotional words explained the memory performance rather than a sensitivity one. The same was true for older adults, but the bias was only towards positive words. Researchers examining this bias have elicited a two-alternative-forced choice paradigm. Here participants make a choice between one target and one foil. Theoretically, when matched for emotional status, bias is removed and the procedure provides a relatively pure measure of accuracy. When pairs are manipulated by emotion and item type (either both target or both foil), bias can be measured toward the emotional item. This same procedure was adopted by Grider and Malmberg who also used foil pairs (e.g. negative or positive and neutral foil pair) to examine bias towards the emotional foil. In this condition, they only found a positive bias effect (more like to choose a positive foil over a neutral foil) but no negative bias effect. These findings differ somewhat from Thapar and Rouder (2009) and Dougal and Rotello (2007), although it was noted that Grider and Malmberg's findings were somewhat modest with a large sample size.

Whilst research continues to examine emotional effects on memory accuracy, another avenue of research has examined the heightened effect of emotion on false memory. For the purposes of this present study, we focus on false memories that are naturally occurring distortions in memory without external suggestive information. They are often referred to as spontaneous false memories (e.g. Brainerd & Reyna, 1998, 2007; Howe et al., 2009). In the laboratory, such false memories have been studied using a prominent procedure known as the Deese/Roediger/Mc- Dermott (DRM; Deese, 1959; Roediger & McDermott, 1995) paradigm. Here, participants study lists of words (*e.g. toe, ankle, shoe, sock, boot, kick*) which are all associated with a critical lure that is not presented with the list (e.g. *foot*). At recall or recognition, if the participants freely recollect the critical lure or recognise it amongst list and distractor items, a false memory is recorded (Roediger & McDermott, 1995).

False memory research has important implications for the forensic field but recollections of events in those circumstances are, by their very nature, affect-laden. One of the most enduring questions about false memory is how it is influenced by emotions that may accompany the past experience (Bookbinder & Brainerd, 2016). One key benefit of the DRM paradigm is its adaptability for testing the effects of emotion. One way to do this is to manipulate valence and arousal in a list stimuli (e.g. *harm*, *pain*, *wound*, *punish*, *insult* [critical lure = hurt]). Budson et al. (2006) were one of the first to demonstrate that emotionally charged keywords could be falsely remembered quite reliably. Since then, numerous studies have adjusted the emotional content of word lists to explore the concept of emotional false memories (Brainerd et al., 2010; Hellenthal et al., 2019; Howe, et al., 2010; Knott et al., 2018; Otgaar et al., 2016). It appears that negative high arousing stimuli provide the optimum conditions for false memory production (Brainerd et al., 2010), however, this is often only evident for false recognition, not recall, with more false recognition responses for emotionally negative compared to neutral critical lures but fewer false recollections in a free recall test for negative compared to neutral critical items (Howe et al.).

Memory-based accounts of false memory have been used to explain this increased emotional false memory effect. Spreading activation models such as the associative activation theory (AAT; Howe et al., 2009) and activation monitoring theories (AMT; Roediger, et al., 2001) posit that when an item is studied, it can activate related but non-presented items in the mental lexicon due to the spreading activation of conceptual representations. The strength of activation of the related but non-presented items increases the difficulty of making diagnostic source monitoring decisions (Roediger & McDermott, 1995) about the presence or absence of that item in the list. As negative emotional information is represented in a dense associative network, it facilitates critical lure activation (Otgaar et al., 2016; Shah & Knott, 2018). Fewer theme nodes mean faster activation spread to the negative critical items. Fuzzy-trace theory is a dual-process account (FTT; Brainerd & Reyna, 1998, 2005). Here, gist traces represent the core meaning of the memory but not its specific details, whereas verbatim traces capture the specific attributes of the memory (e.g. visual features). Retrieving the verbatim traces

results in accurately recognising presented list items. Retrieving gist traces, on the other hand, leads to a feeling of familiarity with the item which can either result in true recognition or false recognition. According to Bookbinder and Brainerd (2016), the presence of negative emotional content in stimuli amplifies the formation of connections between items in the associative list. This, in turn, results in increased levels of false memories by strengthening the overall conceptual or gist-based information while weakening the specific verbatim details associated with the stimuli.

The majority of studies reporting enhanced emotional false memory effects predominantly focus on memory-based explanations. Explanations based on criterion shifts have received relatively less attention. Some previous findings have suggested that response bias for negative critical lures might be more lenient than for neutral critical lures. Howe et al. (2010) argued that higher levels of semantic density and fewer distinct theme nodes, make items seem more familiar and cause greater confusability between what was present or absent in study. In a recognition test, this makes it harder to accurately reject the false lure. Researchers have argued that this increased meaning-based familiarity or confusability causes us to adopt a more lenient criterion for accepting emotional stimuli as old. Indeed, there has been some evidence using signal detection analyses to support this suggestion (see Hellenthal et al., 2019; Yüvrük and Kapucu, 2022).

Criterion shift accounts have in fact been used to explain the DRM's robust findings (Miller & Wolford, 1999; Miller et al., 2011; see Wixted & Stretch, 2000, for a review). The criterion shift perspective asserts that the memory of the critical lure is influenced, to some extent, by a shift in criteria towards a more lenient response compared to other words. That is, participants demand less evidence in the recognition test to accept items that seem familiar to one of the studied themes. Because critical lures are more related to the gist of the study lists, participants adopt a more liberal criterion for critical lure items compared to other items (Miller et al., 2011). This account implies that the recognition decision is not based on the experience of studying the word (falsely through associative activation) but instead on a strategic inferred judgment. In support, Miller and Wolford (1999) found that the measured response criterion for critical items was significantly more liberal than for the other item types. 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.

Other methods have been used to examine the criterion shift account of false memory formation. For example, Miller et al. (2011) designed what they called, a criterion warning, which instructed subjects to avoid responding old to any test items that seem to be related to one of the study lists. The criterion warning significantly reduced the false alarm rate for the critical lure from 77% to 46%, indicating that the false recognition effect may at least be partially explained by a strategic judgement difference. Such a warning was much more effective compared to a typical warning that instructed participants to be aware of highly associated critical lures in the recognition test, presumably because it is not expected that critical lures can be explicitly identified per se. Although Miller et al (2011) found significant decreases in mistakenly recognising related but non-presented items when warnings are given before a memory test, other studies (i.e. Anastasi et al., 2000; Gallo et al., 2001; Neuschatz et al., 2001) have observed minimal or no impact from such warnings. Miller argued that this inconsistency in findings might stem from how effective the warnings are and to what extent they focus on monitoring specific types of items as opposed to generally being aware of the relatedness among items. Indeed, it is this relatedness that may explain the apparent difference in negative and neutral critical lure false alarm rates.

Researchers have also made use of the two-alternative-forced choice paradigm to examine criterion shift. Jou et al. (2018) showed that the rate of false recognition could be greatly reduced when a presented list item was paired with a critical lure. Such a forced choice restricts the role of criterion. This is because subjects do not compare each item against the adopted criterion (whether this be conservative or liberal) as they do in the typical DRM recognition test (yes/no – YN; old/new). Instead, participants compare two items against each other and choose the one with the highest signal value. Jou et al., found that when the list item and the critical lure were presented as a pair, false recognition rates of the critical lure dramatically reduced, although one should note that they are not entirely eliminated. Because using criterion warning and a "criterion-free" test resulted in a significant reduction of false memories, it is difficult to argue against, at least in part, the role of criterion in false recognition

(Jou et al., 2018; Miller et al., 2011), although it is important to emphasise that even in the literature it has been stated that these two possible mechanisms (associative memory and decision processes) are not mutually exclusive, and both may simultaneously contribute to the DRM false memory effect (Miller et al., 2011).

In summary, research has highlighted that emotionally charged stimuli, especially those with negative valence and high arousal, are more susceptible to false memory production. This phenomenon has been explored through various theoretical models like the associative activation theory and fuzzy-trace theory, both emphasising the role of emotional content in enhancing associative networks and gist-based processing, respectively. However, alongside these memory-based explanations, the concept of response bias offers a significant perspective. Studies suggest that a more lenient criterion for accepting emotional stimuli as remembered is influenced by their meaning-based familiarity. This leniency towards emotional stimuli could be the cause of the inherent difficulty in distinguishing between true and false memories.

The utilisation of paradigms such as the two-alternative-forced choice (2AFC) or criterion warnings offers a promising avenue for further research. As suggested by Jou et al. (2018), by constraining the decision process to a comparison between two items, this paradigm minimises the influence of response bias, providing a clearer measure of memory accuracy. This approach is particularly beneficial in understanding the role of emotional content in false memory formation. When applied to emotional false memories, the 2AFC paradigm, and by a similar vain, conservative criterion warning, could help disentangle the contributions of genuine memory distortions from those influenced by decision-making biases. That is, if the enhanced negative false memory effect can be attributed to criterion shifts, it is plausible that employing strategies such as encouraging conservative responding through criterion warnings (Miller et al., 2011) or utilising criterion-free recognition tests (where criterion plays a lesser role compared to a Yes/No recognition test) may eliminate or significantly diminish this effect. However, if an elevated negative false memory effect persists even with more conservative criterion-warning instructions or criterion-free tests, it would suggest that the susceptibility to error formation with negative emotional stimuli is primarily rooted in memory-based explanations. Experiment one will examine the impact of criterion warnings on the production of emotional false memories. We will explore whether a more conservative decision-making approach, prompted by criterion warnings, can effectively reduce the enhanced negative emotional false memory effect. Experiment two will utilise the two-alternative-forced choice task to investigate the role of criterion-free tests on the generation of emotional false memories. Both experiments will use negative arousing vs. neutral non-arousing stimuli in line with Thapar and Rouder (2009) but also emotional false memory studies that find negative arousing DRM lists to produce the largest emotional false memory effect compared to neutral lists (see Brainerd et al., 2010; Knott et al., 2018)

Experiment 1

In Experiment 1, we examined the impact of warning conditions on false recognition for negative and neutral list items. We utilised the warning conditions adopted by Miller et al. (2011) which included a criterion warning and a more standard critical lure warning. According to Miller et al., a criterion warning expressly warns the participants that they should watch out for a word that was related to the theme of the related words and be sure to reject that word because that word would not have been studied. They argued that this differs from a critical lure warning which warns participants of the nature of the task and to avoid responding "old" to critical items. Miller et al., found that in the criterion warning condition, there was a significant reduction in false alarm rates which was associated with a conservative criterion shift. They concluded that at least to a certain extent, this type of warning can impact DRM false recognition if that warning helps participants recognise the general theme or gist causing a criterion shift. To investigate the potential impact of decision-making shifts from false alarms to negative emotional critical lures, we employed three distinct warning conditions: critical Lure warning, a Criterion warning, and a no warning control group. Participants were randomly allocated to one of these conditions.

Method

Participants

One hundred and eighty participants took part in Experiment 1. They were recruited via the online

participant recruitment platform Prolific or were first-year psychology students completing the study for course credit at City, University of London. A priori power analysis indicated a total sample size minimum of 108, with a medium effect size of $f = 0.25$ and Power (a = 0.05, 1 − β err prob) of 0.95. The age range of the sample was 18–60 ($M = 28.49$, $SD = 12.93$) with 76 males. Participants volunteered to take part in the study and were all native-English speakers. Fifty-nine participants were randomly allocated to the No warning group, 60 to the Criterion-warning group, and 61 to the Critical Lure warning group. The mean age across each warning condition did not differ significantly, $F(2, 159) = .72$, $p = .49$, $\eta_p^2 = .01$.

Design and stimuli

The warning condition was a between-participants factor, participants either received no warning, a critical lure warning or a criterion warning. The three warning types were taken from Miller et al. (2011, see the appendix). Emotion was a within-participants factor, all participants were presented with both neutral and negative word lists during the study phase.

DRM lists

We used 16 DRM lists (8 neutral lists and 8 negative lists). The 8 neutral lists were taken from Hellenthal et al. (2019) and Roediger et al. (2001) and had the following neutral critical lures; *car*, *chair*, *smell*, *pen*, *high*, *door*, *foot*, and *mountain*. The 8 emotional-negative lists were taken from Hellenthal et al (2019) and they consisted of the top eight negatively valenced associates in terms of BAS to the following critical lures: *anger*, *cry*, *lie*, *sick*, *hurt*, *thief*, *danger*, and *alone*. All lists contained 12 items. For those items where the values were available, mean valence and arousal ratings for list items and critical lures were taken from the affective norms for English words (ANEW; Bradley & Lang, 1999) database. Independent samples *t*-tests showed that the negative list items (and associated critical lures) had significantly lower ratings of valence but higher ratings of arousal compared to neutral list items (and critical lures). The negative and neutral lists were matched for BAS (see Table 1). We also performed Bayesian independent samples t-tests. Table 1 provides the Bayes Factor (BF_{10}) for each stimulus characteristic and their interpretation following Jeffreys (1961). The outcomes from the Bayes factor analysis were consistent with the results above. That is, there was moderate to extreme evidence for valence and arousal differences across negative and neutral stimuli, but anecdotal evidence in favour of an absence of a difference in BAS across stimulus emotion conditions.

Half the lists were presented with the critical item in position 1 (with the sixth item removed) and half were presented without the critical item, thus acting as the critical lure. All lists were presented together but blocked by emotion. Thus participants saw 8 negative lists (lists 1–4 with the critical item presented, lists 5–8 with the critical lure not presented), followed by the 8 neutral lists (critical item presented and not presented in a similar order). Counterbalancing took place for the emotion presentation order and critical item presented order.

Table 1. Mean (and standard deviations) values, with t-test mean comparisons and Bayes Factor (BF) analysis for valence, arousal and backward associative strength by list emotion.

	Negative lists	Neutral lists	*t*-value	*p*-value	BF_{10}
Experiment 1					
List-item valence	3.09 (.58)	5.46 (.26)	−10.56	<.001	[a]1.53×10^5
Critical lure valence	2.38 (.38)	5.96 (1.03)	−8.68	<.001	[a]6252.90
List-item arousal	5.64 (.92)	4.29 (.60)	3.50	=.004	[b]11.06
Critical lure arousal	6.42 (1.13)	4.73 (1.17)	2.74	=.018	[c]3.49
BAS	.19 (.07)	.22 (.09)	−0.77	=.452	[d]0.52
Experiment 2					
List-item valence	3.09 (.46)	5.39 (.31)	−14.23	<.001	[a]2.84×10^9
Critical lure valence	2.43 (.51)	5.85 (.89)	−10.96	<.001	[a]5.03×10^6
List-item arousal	5.69 (.83)	4.31 (.48)	4.97	<.001	[a]329.52
Critical lure arousal	6.27 (1.15)	4.49 (1.07)	3.66	=.002	[b]20.30
BAS	.21 (.08)	.22 (.07)	−0.32	=.752	[d]0.39

[a]Extreme evidence in favour of a significant difference between Negative and Neutral lists.
[b]Strong evidence in favour of a significant difference between Negative and Neutral lists.
[c]Moderate evidence in favour of a significant difference between Negative and Neutral lists.
[d]Anecdotal evidence in favour of no difference between Negative and Neutral lists.

Recognition test items

The recognition test consisted of 96 items, including 48 presented and 48 non-presented items. The presented list items in the test included the 1st (note, this is the presented critical item for half the lists), 5th, and 10th items from all 16 lists. The non-presented test items consisted of 8 non-presented critical lures, 24 unrelated distractor items, and 16 weakly related distractor items. There are two key requisites of selecting unrelated distractor items when comparing emotion as a design factor (see Hellenthal et al., 2019; Howe et al., 2010; Knott et al., 2018). First, the valence and arousal scores are matched to the presented neutral and neutral list items and there are an equal number of each. Thus, the unrelated distractor items were obtained from the ANEW database (Bradley & Lang, 1999) and 12 were neutral and 12 were negative in valence. Second, unrelated distractors should not be weak associates of any critical lures used in the study phase. Thus, the unrelated distractor items were also carefully selected to ensure they did not appear in the list of associated using the University of South Florida Free Association Norms Database (Nelson et al., 1998). To obtain the weakly related distractor items, we identified the critical lure items in the same database and selected the last item from the list of associates with an associative value of 0.02–0.01. All items were presented randomly to the participants during the recognition test.

Procedure

The experiment was conducted online using Qualtrics. Participants took part in a single study phase in which the eight negative lists and eight neutral lists were presented. Participants were given a 10 s break in between each emotion type block of words. The words were presented in the centre of participants' screens for 1.5 s and were separated by a fixation cross which appeared for one second. Words were presented in black, 72-point, Times New Roman font on a white background. Attention checks were included during the study phase to confirm participants were paying attention to the items being presented to them. Two attention checks were placed randomly in each block of negative and neutral word lists (but not within a running list presentation) and participants had three seconds to click on a button before the page progressed.

After the study phase, participants undertook a five-minute distractor task, which involved completing a series of mathematical problems before moving on to the recognition test. Those in the two warning conditions were also asked to watch a video, which provided either the critical lure warning or the criterion warning verbally. After the warnings were presented, participants completed a comprehension check where they were asked to summarise what they had just heard in their own words. They were informed about this comprehension check before watching the warning video. Meanwhile, participants in the no-warning condition continued with the mathematical problems for a similar duration. Subsequently, all participants were instructed to complete the recognition test. They were required to click "yes" if they believed the word had been presented during the study phase (thus considering it "old") and "no" if they thought the word had not been previously presented (therefore classifying it as "new"). All data for both experiments are available at https://osf.io/xbwzn/. Ethical approval for the experiments was granted by the psychology departmental ethics committee. All participants provided informed consent before taking part, and all procedures were performed in compliance with institutional guidelines.

Results and discussion

Data from 18 participants were removed from the analysis due to failure of the comprehension check for warning instructions[1] or for failing to respond to all of the attention checks (100% failure). Separate ANOVAs were conducted to examine *old response rates* (hit and false alarm rates) followed by discrimination sensitivity and bias. Bonferroni corrected multiple comparisons were used for all significant main effects and interactions. We calculated memory accuracy and response bias measures using discrimination sensitivity (d_a) and bias (c_a). Discrimination sensitivity (d_a) measures the ability to distinguish between old and new items, while c_a measure participants' bias to respond "old" or "new". Higher values of d_a indicate better discrimination (higher memory accuracy) whilst lower values of c_a indicate a higher liberal bias towards the "old" response (Macmillan & Creelman, 2004). We conducted accuracy measures for the discrimination of critical lures from unrelated distractors (d_{a_CL}), and for the discrimination of list items from unrelated distractors ($d_{a\_\text{List items}}$).[2] 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). Here, d_{a_CL} represents false memory performance, whereby higher values indicate higher FMs produced in the DRM paradigm and $d_{a\_List\ items}$ represents standard recognition memory performance whereby higher values indicate higher recognition accuracy. In addition, two bias measures were calculated; c_{a_CL} indicates bias used to discriminate studied critical items from critical lures and $c_{a\_List\ item}$ indicates bias used to discriminate list words from unrelated distractor items.

Old response rates

We conducted separate 2(emotion: negative vs. neutral) × 3(warning condition: no warning vs. critical lure warning vs. criterion warning) mixed factor ANOVAs on each item type with repeated measures on the first factor. There were more negative compared to neutral false alarms to critical lures, $F(1, 159) = 21.59$, $p < .001$, $\eta_p^2 = .12$. There was a significant main effect of the warning condition, $F(2, 159) = 9.84$, $p < .001$, $\eta_p^2 = .11$. Pairwise comparisons showed no significant difference in FA rate between the control and critical lure warning condition ($p = .43$), but there was a significant reduction in FA rate between the criterion-warning condition and control ($p < .001$), and criterion-warning and critical lure warning ($p = .01$). There was no significant interaction, $F(2, 159) = .006$, $p = .99$, $\eta_p^2 = .001$, therefore, across all three warning conditions, negative critical lure FA rate was higher than neutral FA rate (see Figure 1).

For hit rates, critical items that were presented in the list showed no significant difference in emotion $F(1, 159) = .47$, $p = .49$, $\eta_p^2 = .003$, warning condition, $F(2, 159) = 2.93$, $p = .06$, $\eta_p^2 = .04$, and no interaction, $F(2, 159) = .58$, $p = .56$, $\eta_p^2 = .007$. Similarly, the hit rate for studied items showed no significant difference in emotion $F(1, 159) = 1.37$, $p = .24$, $\eta_p^2 = .009$, warning condition, $F(2, 159) = 2.52$, $p = .08$, $\eta_p^2 = .03$, and no interaction, $F(2, 159) = .71$, $p = .49$, $\eta_p^2 = .009$. For weak-related distractors, there were more FAs for negative compared to neutral items, $F(1, 159) = 27.52$, $p < .001$, $\eta_p^2 = .15$, but there was no main effect of warning condition, $F(2, 159) = .10$, $p = .91$, $\eta_p^2 = .001$, and no interaction, $F(2, 159) = 2.12$, $p = .12$, $\eta_p^2 = .03$. Similarly for unrelated distractors, there were more FAs for negative compared to neutral items, $F(1, 159) = 19.93$, $p < .001$, $\eta_p^2 = .11$, but there was no main effect of warning condition, $F(2, 159) = .02$, $p = .98$, $\eta_p^2 = .001$, and no interaction, $F(2, 159) = 1.43$, $p = .24$, $\eta_p^2 = .02$ (see Table 2).

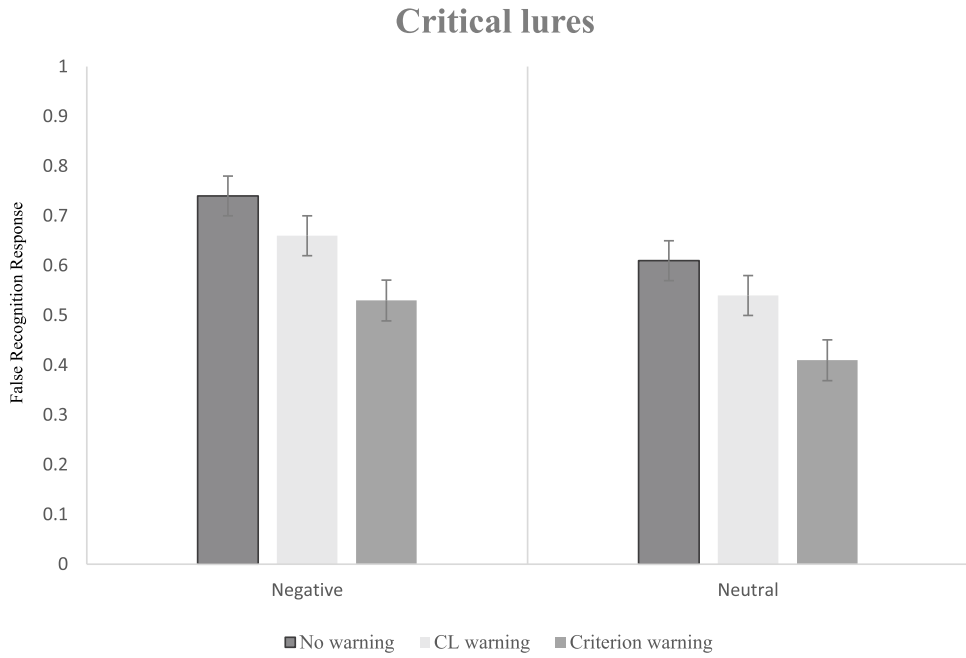

Figure 1. Proportion of false alarms to critical lures as a function of emotion and warning condition for Experiment 1(error bars represent standard error).

Table 2. Mean Hits and FAs (and standard deviation) values as a function of warning condition and emotion for Experiment 1.

	Negative Items			Neutral Items		
	No warning	Critical Lure warning	Criterion warning	No warning	Critical Lure warning	Criterion warning
List word	.64 (.19)	.63 (.21)	.55 (.19)	.62 (.21)	.59 (.21)	.56 (.22)
Studied critical item	.81 (.27)	.75 (.28)	.67 (.28)	.75 (.27)	.74 (.27)	.69 (.30)
Critical lure	.74 (.30)	.67 (.28)	.53 (.30)	.61 (.30)	.54 (.32)	.41 (.29)
Related distractor	.32 (.21)	.28 (.18)	.26 (.24)	.18 (.22)	.19 (.21)	.21 (.22)
Unrelated distractor	.18 (.17)	.16 (.15)	.17 (.16)	.11 (.15)	.13 (.15)	.12 (.17)

Memory sensitivity for d_{a_CL} and $d_{a\_List\ Item}$

Similar 2(emotion: negative vs. neutral) × 3(warning condition: no warning vs. critical lure warning vs. criterion warning) mixed ANOVAs were conducted for each sensitivity measure. For d_{a_CL}, there was no significant main effect of emotion, $F(1, 159) = 1.91$, $p = .17$, $\eta_p^2 = .01$, nor significant interaction, $F(2, 159) = .32$, $p = .73$, $\eta_p^2 = .004$ but there was a significant main effect of warning condition, $F(2, 159) = 7.57$, $p < .001$, $\eta_p^2 = .09$. Pairwise comparisons showed no significant difference in sensitivity to critical lures between the no warning ($M = 1.54$, $SE = .09$) and critical lure warning conditions ($M = 1.31$, $SE = .09$, $p = .20$), but d_a was lower in the criterion warning condition compared to no warning ($M = 1.54$, $SE = .09$, $p < .001$), indicating fewer FAs to critical lures. There was no significant difference between criterion and critical lure warning ($p = .13$). A similar analysis for $d_{a\_List\ Item}$ revealed a significant main effect of emotion, $F(1, 159) = 8.02$, $p = .005$, $\eta_p^2 = .05$, with better memory accuracy for neutral vs. negative list items. However, there was no significant main effect of warning, $F(2, 159) = 1.11$, $p = .33$, $\eta_p^2 = .01$ or interaction, $F(2, 159) = 1.87$, $p = .16$, $\eta_p^2 = .02$.

Response bias for c_{a_CL} and $c_{a\_List\ Item}$

Analysis of response bias, indicated more liberal responding for negative versus neutral items, $F(1, 159) = 40.28$, $p < .001$, $\eta_p^2 = .20$. Although there was no significant interaction, $F(2, 159) = .33$, $p = .72$, $\eta_p^2 = .004$, there was a main of warning condition, $F(2, 161) = 4.03$, $p = .02$, $\eta_p^2 = .05$. The criterion warning produced the most conservative response bias ($M = .59$, $SE = .06$), which was significantly higher than the no-warning condition ($M = .34$, $SE = .06$, $p = .02$), although not significantly higher than the critical lure warning ($M = .41$, $SE = .06$, $p = .16$). There was no significant difference in response bias for the no warning and critical lure warning conditions ($p = 1$). Although the criterion did decrease for negative critical lures in the criterion-warning condition, it was still more liberal than the criterion value for neutral critical lures (see Table 1). For $c_{a\_List\ Item}$ measures, there was also a main effect of emotion, $F(1, 159) = 14.59$, $p < .001$, $\eta_p^2 = .08$, with a more liberal response bias for negative vs. neutral items. There was no significant main effect of warning, $F(2, 159) = 1.55$, $p = .22$, $\eta_p^2 = .02$ or interaction, $F(2, 159) = .36$, $p = .70$, $\eta_p^2 = .005$. Warning does not impact bias in decision-making for list items which is in line with our expectations (see Table 3).

We acknowledge that Miller et al.'s criterion warning has been criticised for its effectiveness in shifting criterion setting at retrieval because it assumes that criterion setting and/or identifying the critical lure is a conscious process. Jou et al. (2018) argued instead that a criterion shift explanation should be tested by using a recognition test that is considered to be criterion free (e.g. Hicks & Marsh, 1998; Macmillan & Creelman, 2004). This will be the focus of Experiment 2.

Table 3. Mean d_a and C_a (and standard deviation) values as a function of warning condition and emotion for Experiment 1.

	Negative items			Neutral items		
	No warning	Critical lure warning	Criterion warning	No warning	Critical lure warning	Criterion warning
d_{a_CL}	1.56 (.84)	1.40 (.68)	1.09 (.89)	1.52 (.91)	1.22 (.86)	1.01 (.69)
C_{a_CL}	.19 (.54)	.30 (.51)	.46 (.56)	.49 (.47)	.52 (.53)	.72 (.57)
$d_{a\_List\ Item}$	1.37 (.84)	1.39 (.77)	1.15 (.75)	1.59 (.92)	1.39 (.81)	1.39 (.85)
$C_{a\_List\ Item}$	.29 (.41)	.31 (.44)	.43 (.44)	.45 (.37)	.43 (.43)	.53 (.46)

Experiment 2

In Experiment 2, we adopted a procedure similar to Experiment 4 in Jou et al. (2018). Researchers have argued that a two-alternative-forced choice test (2AFC) is considered to be a criterion-free test (e.g. Hicks & Marsh, 1998; Macmillan & Creelman, 2004) or at least considerably reduced (Jou et al., 2018) in comparison to a yes/no (YN) test. We therefore compared rates of false alarms for negative and neutral critical lures across different test conditions: Yes/No recognition (YN) test, two-alternative-forced choice (2AFC) test, and two-alternative free-choice test (2AFrC). We included this latter test to determine whether the decrease in false-recognition rate of critical lures in the 2AFC test, compared to the YN test, was due to a restricted criterion role in the 2AFC or because the probe pair mate in the 2AFC provides helpful clues for identifying the target item. A 2AFrC test is in effect equivalent to presenting two Yes/No test items simultaneously in one test trial. This is important because the requirement of having to choose one and only one item in a 2AFC is removed. Participants can choose either or both items in the pair, therefore they can resume the adoption of an absolute criterion in the 2AFRC test if they so desire. This test condition will help determine whether it is the decision criterion per se or some information afforded by a pair mate that causes a lowered critical lure FA rate in a 2AFC condition. So, in Experiment 2, participants were allocated to one of the three test conditions. Each participant was presented with 6 DRM lists before being tested. The lists were blocked by emotion so that recognition tests were all negative or all neutral stimuli. This method differed from Experiment 1 where all negative and neutral lists were presented in one study block followed by the recognition test. This was required to ensure that once the participants had heard the warning instruction at the test, we did not confound further encoding phases from potential effects of the previous warning instruction. In Experiment 2, we utilised the study/test blocks as used by Jou et al. This also allowed us to have separate recognition tests for negative and neutral items. A point we will refer to in the general discussion.

Method

Participants

One hundred and fifty-five participants completed the online study in return for a small fee. The age range of the participants was 18–59 ($M = 34.98$, $SD = 12.84$ with 52 males). There were 50 participants in the Yes/No (YN) test condition, 54 in the two-alternative-forced choice (2AFC) condition, and 51 in the two-alternative free choices (2AFrC). The mean age across each warning condition did not differ significantly, $F(2, 145) = .24$, $p = .78$, $\eta_p^2 = .003$. All participants were native-English speakers of the UK nationality. A priori power analysis indicated a total sample size of 108, with a medium effect size of $f = 0.25$ and Power ($a = 0.05$, $1 - \beta$ err prob) of 0.95. Informed consent was obtained from all participants and they were debriefed at the end of the experiment.

Design and stimuli

We compared three test conditions, the standard 2AFC, YN, and 2AFrC. Test format was a between-participants condition and emotion (neutral vs. negative word lists) was a within-participants condition.

DRM lists

For this experiment, we used twenty-four of the DRM lists (12 neutral and 12 emotional-negative) taken from Hellenthal et al. (2019) and Roediger et al. (2001) and this time consisted of the top twelve associates in terms of Backward Associate Strength (BAS) to the following neutral critical lures: *car*, *chair*, *smell*, *pen*, *high*, *door*, *foot*, *mountain*, *window*, *shirt*, *cup*, and *eye* and negative critical lures: *alone*, *anger*, *dead*, *gun*, *sick*, *thief*, *cry*, *hate*, *lie*, *danger*, *hurt*, and *fear*. Independent Samples t-tests showed that the negative list items (and associated critical lures) had significantly lower ratings of valence but higher ratings of arousal compared to neutral list items (and critical lures). The negative and neutral lists did not significantly differ in BAS (see Table 1). Similar to Experiment 1, we performed Bayesian independent samples t-tests. Table 1 provides the Bayes Factor (BF_{10}) for each stimulus characteristic and their interpretation. Once again, the outcomes from the Bayes factor analysis were consistent with the results above. That is, there was strong to extreme evidence for valence and arousal differences across negative and neutral stimuli, but anecdotal evidence in favour of an absence of a difference in BAS across stimulus emotion conditions.

Study/Test trial blocks. Each Participant completed 4 study/test trial blocks. The blocks were separated by emotion so that two study/test trials used neutral

lists and two used negative lists. Within each block, three lists included the critical item (e.g. *car*, *chair*, *pen*) as the first word of the list and three did not (standard list with critical lure not presented). For lists that included the critical item in position one, item six of that list was removed. For each study/test phase, there were 6 lists followed by a 36-item recognition test (18 presented and 18 non-presented items). The items in the recognition test included three words from each list (from the 1st, 5th and 10th position [note that position 1 from the critical presented lists is the critical item]). There were three non-presented critical lures, three related distractors (the removed item from position 6)[3] and 12 unrelated distractors (6 neutral 6 negative, matched by valence and arousal using ANEW values). Full counterbalancing took place regarding the order of the blocks and the use of lists as presented or non-presented critical items lists.

Procedure

Study phase

The entire experiment was conducted online each word was presented centrally on the screen for 1.5s with a 1-s inter-stimulus interval. Words were presented in black, 72-point, Times New Roman font on a white background. Each item was presented in descending order of BAS. A 5-minute distractor task (a Sudoku or Maze puzzle) preceded the recognition test of each block.

Test phase

For the YN test condition each of the 36 items was presented on screen in random order, Participants were asked to make a yes decision, by clicking on the yes button, if they recognised the item from the study phase and a no-decision, by clicking on the no button, if they did not recognise the word from the study phase. In the 2AFC condition, each of the three presented target words from a list (the three list words from Positions 1, 5, and 10) were randomly paired with a non-presented item. For lists that did not include the critical item, the three non-presented items were two unrelated words and the critical lure. For example, Chair – Couch. For lists that did include the critical item, the three non-presented items were two unrelated words plus the 6th item which was removed from the list presented. The target and distractor were assigned to the left and right sides of the screen. For each pair, the location to the left or right occurred with equal probability, and target and distractor sides were counterbalanced across subjects. Participants were told that one of the words was from the list and they had to click on the left or right button (underneath each word) to signal which was old. The 2AFrC condition followed the same presentation of items as the 2AFC condition but a change to the instruction given to the participants. Here, participants were not told that one of the words in the pair was from the list and one was not. Instead, they were just instructed to click the "both" button if they recognised both items, click the "neither" button if did not remember either word, or click the left or right button if they remembered studying the left-side word, or the right-side word. Finally, throughout each block, two attention checks were placed randomly (but not within the presentation of a list) and participants had three seconds to click on a button before the page progressed.

Results and discussion

Data from seven participants were removed for failing to respond to all of the attention checks (100% failure). Note that hits for studied list words were calculated from items chosen when paired with a distractor (for 2AFC and 2AFrC). We used Bonferroni-corrected multiple comparisons for all significant main effects and interactions. Like Experiment 1, we conducted separate ANOVAs to examine *old response rates* (hit and FA rates) followed by discrimination sensitivity and bias. Again, we conducted accuracy measures for the discrimination of critical lures from unrelated distractors (d_{a_CL}), and for the discrimination of list items from unrelated distractors ($d_{a\_\text{List items}}$). In addition, two bias measures were calculated; c_{a_CL} indicates bias used to discriminate studied critical items from critical lures and $c_{a\_\text{List item}}$ indicates bias used to discriminate list words from unrelated distractor items (see Table 5).

Old response rates

We conducted separate 2(emotion: negative vs. neutral) × 3(test format: YN vs. 2AFC vs. 2AFrC) mixed ANOVAs on each item type with emotion as the repeated measures factor. For false alarm rates to critical lures, there were significant main effects of emotion, $F(1, 145) = 40.10$, $p < .001$, $\eta_p^2 = .22$ and test format, $F(2, 145) = 48.08$, $p < .001$, $\eta_p^2 = .40$. Although there was a significant emotion × test

format interaction, $F(2, 145) = 7.12$, $p < .001$, $\eta_p^2 = .09$. We conducted separate one-way ANOVAs for false alarms to negative and neutral critical lures across the three test formats. For both negative critical lures, $F(2,145) = 28.74$, $p < .001$, $\eta_p^2 = .28$ and neutral critical lures, $F(2,145) = 48.77$, $p < .001$, $\eta_p^2 = .40$, false response rates decreased significantly in the 2AFC compared to YN and 2AFrC conditions ($ps < .001$), with no difference in the latter two conditions ($ps = 1.00$). Although Figure 2 shows that this drop in false alarms in the 2AFC format appears greater for neutral critical lures, there was no statistical difference. Comparison of emotion at each test condition indicated greater false alarm rates to negative compared to neutral critical lures in the 2AFrC ($p = .03$), YN ($p = .05$), and 2AFC ($p < .001$), however the magnitude of difference in the final 2AFC condition was higher.

The hit rate for presented critical words showed no difference in test format, $F(2, 145) = 0.24$, $p = .79$, $\eta_p^2 = .003$ but more hits for negative vs. neutral presented critical items, $F(1, 145) = 7.35$, $p = .008$, $\eta_p^2 = .05$. There was no significant interaction, $F(2, 145) = 2.61$, $p = .08$, $\eta_p^2 = .04$. Hit rate for studied items showed no difference in test format, emotion and no significant interaction (all $p > .05$). For related distractors, there was a main effect of test type, $F(2, 145) = 10.37$, $p < .001$, $\eta_p^2 = .13$, with more FAs in the YN and 2AFrC tasks compared to 2AFC (both $ps < .05$), but there was only a marginal effect of emotion, $F(1, 145) = 3.87$, $p = .051$, $\eta_p^2 = .03$ and no interaction, $F(2, 145) = 1.25$, $p = .29$, $\eta_p^2 = .02$. There were no significant effects for FAs to unrelated distractors (all $p > .05$).

Additionally, we compared old decisions for critical lures versus studied list items when presented as a pair in the 2AFC and 2AFrC conditions. By doing so, we could examine the relative memory activation for negative and neutral critical lures when they were paired with an equivalent emotional list item. For negative items, there was no significant difference between the two item types for 2AFrC ($M = .81$ vs. $M = .80$, respectively) and 2AFC ($M = .51$ vs. $M = .49$, $ps > .79$). However, for neutral items, there were significantly more hits to studied words compared to FAs for critical lures in both 2AFrC ($M = .88$ vs. $M = .73$, $p = .002$) and 2AFC conditions ($M = .72$ vs. $M = .28$, $p < .001$). Jou et al. (2018) conducted this same analysis, with a similar outcome whereby a much larger reduction in critical lure FAs was seen in the 2AFC condition compared to 2AFrC. They argued that pairing a studied word with the critical lure was not enough to lead participants to dismiss the critical lure (i.e. the 2AFrC condition) and that the large drop in critical lure FA rate in the 2AFC condition was due to the reduced reliance on an absolute

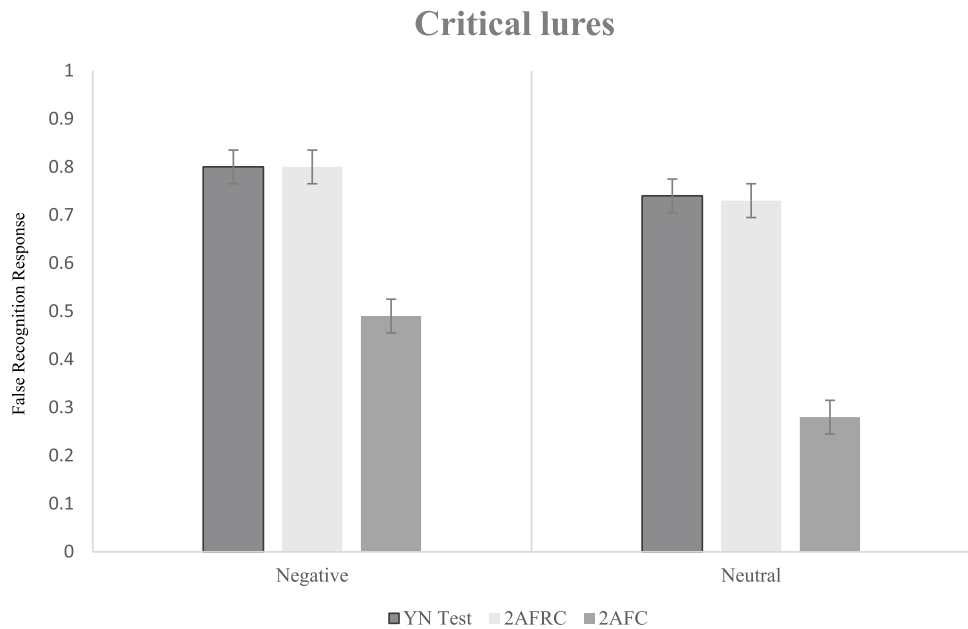

Figure 2. Proportion of false alarms to critical lures as a function of emotion and test type for Experiment 2(error bars represent standard error).

Table 4. Mean old responses (and standard deviation) as a function of test type, item type and emotion for Experiment 2.

	Negative items			Neutral items		
	YN	2AFC	2AFrC	YN	2AFC	2AfrC
Item type						
List word	.74 (.16)	.79 (.13)	.78 (.11)	.77 (.17)	.81 (.13)	.78 (.14)
List word (paired with critical lure)		.51 (.21)	.81 (.15)		.72 (.19)	.87 (.14)
List word (paired with unrelated distractor)		.88 (.15)	.77 (.13)		.87 (.13)	.76 (.15)
Studied critical item	.91 (.15)	.95 (.09)	.95 (.09)	.92 (.18)	.90 (.15)	.88 (.14)
Critical lure	.80 (.24)	.49 (.21)	.80 (.25)	.74 (.30)	.28 (.19)	.73 (.27)
Related distractor	.37 (.22)	.20 (.16)	.33 (.21)	.37 (.26)	.24 (.22)	.42 (.30)
Unrelated distractor	.09 (.12)	.10 (.14)	.07 (.13)	.08 (.11)	.13 (.12)	.10 (.13)

Note: 2AFC, two-alternative-forced choice; 2AfrC, two-alternative free choice.
Values here show the chosen studied word when paired with a unrelated distractor or critical lure for the two-alternative-forced choice and two-alternative free choice.

criterion decision about the critical lure, as opposed to any helpful discrimination that the pair mate made. However, of interest, this did not appear to be the case for the negative emotion condition. Instead, we found no significant difference between hits and FAs in the forced choice task. This suggests an inability to discriminate between presented list items and non-presented critical lures for negative items which led to an almost 50/50 split in the forced choice of either the negative critical lure or the negative presented list item. When forced, choosing the critical lure 28% of the time in the neutral condition vs. 49% of the time in the negative condition demonstrates a stark difference in these two list types, something that was highlighted above in the false alarms analysis. A similar analysis was conducted comparing hits to studied items and FAs to unrelated distractors, but for both test conditions and both emotion types, hits for list items were higher than false alarms to distractor items (all *ps* < .05, see Table 4).

Memory sensitivity for d_{a_CL} and $d_{a\_List\ items}$

We utilised 2(emotion: negative vs. neutral) × 3(test format: YN vs. 2AFC vs. 2AFrC) mixed factor ANOVAs for each sensitivity measure (d_{a_CL}) and ($d_{a\_list\ items}$). For d_{a_CL}, there was a significant main effect of emotion, $F(1, 145) = 14.78$, $p = .008$, $\eta_p^2 = .09$ and test format, $F(2, 145) = 61.51$, $p < .001$, $\eta_p^2 = .46$, which were qualified by a significant interaction, $F(2, 145) = 3.91$, $p = .02$, $\eta_p^2 = .05$. Comparison of emotion at each test condition indicated better memory sensitivity (fewer FAs to critical lures) for neutral compared to negative items in the 2AFC test ($M = 0.73$ vs. $M = 1.40$, $p < .001$), but not for the YN test ($M = 2.10$ vs. $M = 2.25$, $p = .35$), or the 2AFrC test ($M = 2.07$ vs. $M = 2.25$, $p = .76$). For $d_{a\_List\ items}$, there was a significant main effect of emotion, $F(1, 145) = 15.14$, $p < .001$, $\eta_p^2 = .10$ but not test format, $F(2, 145) = .07$, $p = .94$, $\eta_p^2 = .001$. The interaction was significant, $F(2, 145) = 3.20$, $p = .04$, $\eta_p^2 = .04$, which suggested that memory sensitivity for list items was significantly better for neutral versus negative emotion in the YN test format only ($p < .001$, although note, the pattern was in the same direction for 2AFC and 2AFrC).

Response bias for c_{a_CL} and $c_{a\_List\ items}$

Analysis of response bias for critical lures revealed no significant effect of emotion, $F(1, 145) = 1.89$, $p = .17$, $\eta_p^2 = .01$. There was a test format main effect, $F(2, 145) = 26.17$, $p < .001$, $\eta_p^2 = .27$, whereby participants were more liberal in the YN and 2AFrC conditions compared to the 2AFC condition (both *ps* < .001). There was no interaction, $F(2, 145) = .48$, $p = .62$, $\eta_p^2 = .01$. Response bias for list items also showed no main effect of emotion, $F(1, 145) = .36$, $p = .55$, $\eta_p^2 = .002$. There was a significant main effect of test type, $F(2, 145) = 3.52$, $p = .03$, $\eta_p^2 = .05$, but no interaction, $F(2, 145) = .13$, $p = .88$, $\eta_p^2 = .002$ (see Table 4). For test type, there as only one significant

Table 5. Mean d_a and C_a (and standard deviation) values as a function of test type and emotion for Experiment 2.

	Negative items			Neutral items		
	YN Test	2AFC	2AFrC	YN Test	2AFC	2AFrC
d_{a_CL}	2.25 (.85)	1.40 (.63)	2.25 (.74)	2.10 (.96)	.73 (.64)	2.07 (.82)
C_{a_CL}	.33(.47)	.76 (.43)	.37 (.51)	.42 (.54)	.90 (.50)	.38 (.54)
$d_{a\_List\ Items}$	2.16 (.86)	2.37 (.90)	2.29 (.67)	2.68 (.83)	2.48 (.94)	2.46 (.86)
$C_{a\_List\ Items}$	.16 (.30)	.04 (.28)	.12 (.28)	.16 (.34)	.07 (.23)	.16 (.32)

comparison between 2AFC compared to YN condition, $p = .04$ (all other comparisons above $ps > .05$). However, this is likely a result of fewer hits to list items, by its design, in the forced choice condition.

General discussion

The primary objective of this study was to investigate the contributions of genuine memory distortions from those influenced by decision-making biases when eliciting emotional false memories using the DRM paradigm. We utilised two methods previously shown to either minimise the influence of response bias (forced choice test) or shift response biases to be more conservative (criterion warning). If the increased occurrence of negative emotional false memories is due to response biases, then if we compare tactics that promote cautious responding or reduce reliance on criteria compared to standard test conditions (unlike Yes/No tests), might reduce or even eliminate this effect. However, if an elevated negative false memory effect persists it would imply that the tendency to create errors with emotionally negative content is primarily rooted in memory-based explanations.

Experiment 1 showed that warning participants to be careful to respond "old" to any item that was related to one of the studied themes reduced the false alarm rate to critical lures compared to other warning conditions or no warning at all. This effect was similar for false alarms to both negative and neutral critical lure items. However, negative false memories were still higher than neutral false memories in the criterion-warning condition. The relative persistence of false memories despite instructions to shift criterion response, suggests to some extent that a memory-based explanation may be involved. Despite a large reduction in false alarms, in neither case did the false alarm rate to critical lures drop to the levels of false alarm rates for unrelated items. Rather, this reduction may represent the extent to which participants can consciously utilise information about the gist of the study lists to judge whether a test item is old or new. Whilst gist strengthens, the familiarity of critical items it is also convincing us to accept less evidence to decide whether an item is old (Brainerd & Reyna, 1998). If gist extraction is easier for negative versus neutral lists, then familiarity and thus bias will be greater for negative versus neutral critical lures (Bookbinder & Brainerd, 2016). When participants are warned not to rely on the gist, they have fewer errors; however, the negative emotion false alarm effect is still apparent.

To our knowledge, Experiment 2 is the first to use a 2AFC as a means to manipulate and examine the role of criterion in the production of negative and neutral DRM lists. Here, false alarm rates were higher overall for negative critical lures and also in the two-alternative free choice and yes/no recognition tests compared to the two-alternative-forced choice test. This was the case for both negative and neutral critical lures. Participants were also more liberal in the YN and 2AFrC conditions compared to the 2AFC condition but this time, the effect of emotion, did not reach significance. Further, we examined 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. This cannot be achieved by purely examining a 2AFC and a YN recognition test design (Kroll et al., 2002; McKenzie et al., 2001; Smith & Duncan, 2004) but instead requires a free choice condition. Using a similar methodology to Jou et al. (2018), we found that presenting a studied word next to a critical lure in a test pair during a 2AFC condition significantly increased the rejection of the critical lure in favour of the correct hit for the list item (.28 vs .72). In contrast, for negative items, false alarms for critical lures and hit rates for list words did not differ (.49 and .51 respectively). In the free choice pair, critical lures and list items were often both selected, for both neutral and negative, replicating similar findings to the YN condition. Therefore, whenever they were allowed to, participants would accept the critical lure as studied. For neutral false alarms, these findings are in line with Jou et al (2018), but the lack of differentiation between negative list items and critical lures differs. Jou et al. (2018) argued that the critical lure, as a highly distracting item, is memory-based, whereas its transformation from a super-distractor to the status of a studied word is likely criterion-based. If we follow the logic of this explanation for negative items, if there was a similar level of activation in the 2AFC for both negative studied items and critical lures then this suggests that without criterion strategies, memory activation is of a similar level for negative studied items and critical lures. Therefore, memory activation is higher for negative critical lures than it is for neutral critical lures and the difference between emotional false memories is not just a result of more liberal criterion, although it appears that the familiarity and

relatedness of emotional list items increase the readiness with which we are likely to respond old to related but not presented items.

It should be noted that we found within test criterion shifts for both mixed emotion tests and separate emotion test conditions. Note that in Experiment 1, both negative and neutral lists were presented, followed by the recognition test, which presented test items mixed with emotion. We used this method due to the warning condition that could not be replicated in a within participant's emotion condition. In Experiment 2, due to the replication design adapted from Jou et al. (2018), we had separate encoding-test conditions for each emotion type. Stretch and Wixted (1998) argued that trial-by-trial criterion shifts for items differentially encoded (i.e. neutral vs. negative) but tested together does not occur. However, this did not appear to be the case for our two Experiments. We evidenced criterion shifts for negative and neutral items on a trial-by-trial basis for both mixed (although blocked at encoding) and pure list manipulations. Indeed, there have been mixed views and findings on trial-by-trial criterion shifts (for a review see Starns & Olchowski, 2015). Furthermore, an examination of response bias within negative and neutral false decisions has also seen a similar trial-by-trial shift when negative and neutral items were presented together at test (Yüvrük and Kapucu, 2022). Our findings also suggest that this is possible, especially for differing emotional items.

In conclusion, this study investigated the impact of criterion setting on false recognition, specifically focusing on negative emotional valence. The findings support the notion that response bias, influenced by criterion setting, plays a role in the enhanced false recognition effect for negative stimuli. However, even under criterion-free test conditions the heightened emotional false memory effect persisted, and activation of critical lure items appeared to be at a similar level to list items. These results suggest that factors beyond a lenient criterion, which may include implicit associative activation, heightened familiarity, and sensory details contribute to the increased false recognition of negative emotional items. We should note that false memories can be measured more subjectively (by using standard recognition tests or phenomenological reports) or more objectively (by using reaction time measures or indeed, a 2AFC test). If false memories are "contaminated" by subjective judgement, that is, setting a criterion by which you will accept an item as old, then it is apparent from these findings, that this "contamination" appears greater for negative vs. neutral false memories. What we have learnt here is that even with more objective measures of memory, negative emotional items evoke greater false recognition.

Notes

1. Warning instruction comprehension text was independently reviewed by two of the experimenters and only those that reached agreement of failure were removed. In addition, for both Experiments, although data was removed for failing attention checks or comprehension, all other data were included. No data cleaning took place to remove too fast or slow responses. Eyeballing the data did not reveal any potential concerns regarding the need to do this.
2. Note that whilst we used critical lures as the hits to calculate false-recognition sensitivity and bias, typical when using the DRM paradigm (e.g. Arndt & Hirshman, 1998), others, including Jou et al. (2018) have instead suggested that a hit rate of the critical word must be available, as well as its FA rate, hence why we included critical items studied. Here, sensitivity and bias is calculated with the presented critical item as the hit and the non-presented critical lure as the FA rate. Given that we are replicating some of Jou et al.'s methodology, we also conducted similar analyses. This can be found in the supplementary material.
3. In Experiment 2, the related distractor was the nonpresented sixth item of the list. To ensure there was no confounding variable associated with BAS strength for the related distractor, we conducted independent samples *t*-test and found no significant difference across neutral and emotion (.09 and .12, respectively) items, $t(22) = .91$, $p = .37$.

Disclosure statement

No potential conflict of interest was reported by the author(s).

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), 1–26. https://doi.org/10.2307/1423458

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371–391. https://doi.org/10.1006/jmla.1998.2581

Bookbinder, S. H., & Brainerd, C. J. (2016). Emotion and false memory: The context–content paradox. *Psychological Bulletin*, *142*(12), 1315–1351. https://doi.org/10.1037/bul0000077

Bradley, M. M., & Lang, P. J. (1999). Fearfulness and affective evaluations of pictures. *Motivation and Emotion*, *23*(1), 1–13. https://doi.org/10.1023/A:1021375216854

Brainerd, C. J., Holliday, R. E., Reyna, V. F., Yang, Y., & Toglia, M. P. (2010). Developmental reversals in false memory: Effects of emotional valence and arousal. *Journal of Experimental Child Psychology*, *107*(2), 137–154. https://doi.org/10.1016/j.jecp.2010.04.013

Brainerd, C. J., & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, *71*(2), 81–129. https://doi.org/10.1006/jecp.1998.2464

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Brainerd, C. J., & Reyna, V. F. (2007). Explaining developmental reversals in false memory. *Psychological Science*, *18*(5), 442–448. https://doi.org/10.1111/j.1467-9280.2007.01919.x

Budson, A. E., Todman, R. W., Chong, H., Adams, E. H., Kensinger, E. A., Krangel, T. S., & Wright, C. I. (2006). False recognition of emotional word lists in aging and Alzheimer disease. *Cognitive and Behavioral Neurology*, *19*(2), 71–78. https://doi.org/10.1097/01.wnn.0000213905.49525.d0

Cahill, L., Babinsky, R., Markowitsch, H. J., & McGaugh, J. L. (1995). The amygdala and emotional memory. *Nature*, *377*(6547), 295–296. https://doi.org/10.1038/377295a0

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294–299. https://doi.org/10.1016/S0166-2236(97)01214-9

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, *14*(3), 423–429. https://doi.org/10.3758/BF03194083

Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review*, *8*(3), 579–586. https://doi.org/10.3758/BF03196194

Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition*, *36*(5), 933–946. https://doi.org/10.3758/MC.36.5.933

Hellenthal, M. V., Knott, L., Howe, M. L., Wilkinson, S., & Shah, D. (2019). The effects of arousal and attention on emotional false memory formation. *Journal of Memory and Language*, *107*, 54–68. https://doi.org/10.1016/j.jml.2019.03.010

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*(5), 1105–1120. https://doi.org/10.1037/0278-7393.24.5.1105

Howe, M. L., Candel, I., Otgaar, H., Malone, C., & Wimmer, M. C. (2010). Valence and the development of immediate and long-term false memory illusions. *Memory (Hove, England)*, *18*(1), 58–75. https://doi.org/10.1080/09658210903476514

Howe, M. L., Wimmer, M. C., Gagnon, N., & Plumpton, S. (2009). An associative-activation theory of children's and adults' memory illusions. *Journal of Memory and Language*, *60*(2), 229–251. https://doi.org/10.1016/j.jml.2008.10.002

Jeffreys, H. (1961). *Theory of probability (3rd edn)*. Oxford University Press.

Jou, J., Escamilla, E. E., Arredondo, M. L., Pena, L., Zuniga, R., Perez, M., & Garcia, C. (2018). 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. *Quarterly Journal of Experimental Psychology*, *71*(2), 499–521. https://doi.org/10.1080/17470218.2016.1256416

Knott, L. M., Howe, M. L., Toffalini, E., Shah, D., & Humphreys, L. (2018). The role of attention in immediate emotional false memory enhancement. *Emotion*, *18*(8), 1063–1077. https://doi.org/10.1037/emo0000407

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*(2), 241–254. https://doi.org/10.1037/0096-3445.131.2.241

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64. https://doi.org/10.1038/nrn1825

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Cambridge University Press.

Mather, M., & Nesmith, K. (2008). Arousal-enhanced location memory for pictures. *Journal of Memory and Language*, *58*(2), 449–464. https://doi.org/10.1016/j.jml.2007.01.004

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*(1), 140–155. https://doi.org/10.1037/0096-3445.130.1.140

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*(5), 1228–1235. https://doi.org/10.1037/a0024539

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, *106*(2), 398–405. https://doi.org/10.1037/0033-295X.106.2.398

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

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 (Hove, England)*, *9*(1), 53–71. https://doi.org/10.1080/09658210042000076

Otgaar, H., Howe, M. L., Brackmann, N., & Smeets, T. (2016). The malleability of developmental trends in neutral and negative memory illusions. *Journal of Experimental Psychology: General*, *145*(1), 31–55. https://doi.org/10.1037/xge0000127

Roediger, H. L., III & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. https://doi.org/10.3758/BF03196177

Schümann, D., Bayer, J., Talmi, D., & Sommer, T. (2018). Dissociation of immediate and delayed effects of emotional

arousal on episodic memory. *Neurobiology of Learning and Memory*, *148*, 11–19. https://doi.org/10.1016/j.nlm.2017.12.007

Shah, D., & Knott, L. M. (2018). The role of attention at retrieval on the false recognition of negative emotional DRM lists. *Memory* (*Hove*, *England*), *26*(2), 269–276. https://doi.org/10.1080/09658211.2017.1349803

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*(3), 615–625. https://doi.org/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*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*, *43*(1), 49–59. https://doi.org/10.3758/s13421-014-0433-y

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*(6), 1379–1396. https://doi.org/10.1037/0278-7393.24.6.1379

Talmi, D. (2013). Enhanced emotional memory. *Current Directions in Psychological Science*, *22*(6), 430–436. https://doi.org/10.1177/0963721413498893

Talmi, D., Luk, B. T. C., McGarry, L. M., & Moscovitch, M. (2007b). The contribution of relatedness and distinctiveness to emotionally-enhanced memory. *Journal of Memory and Language*, *56*(4), 555–574. https://doi.org/10.1016/j.jml.2007.01.002

Talmi, D., Schimmack, U., Paterson, T., & Moscovitch, M. (2007a). The role of attention and relatedness in emotionally enhanced memory. *Emotion*, *7*(1), 89–102. https://doi.org/10.1037/1528-3542.7.1.89

Thapar, A., & Rouder, J. N. (2009). Aging and recognition memory for emotional words: A bias account. *Psychonomic Bulletin & Review*, *16*(4), 699–704. https://doi.org/10.3758/PBR.16.4.699

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585–594. https://doi.org/10.1016/j.tics.2005.10.011

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory.. *Psychological Review*, *107*(2), 368–376. https://doi.org/10.1037/0033-295X.107.2.368

Yüvrük, E., & Kapucu, A. (2022). False (or biased) memory: Emotion and working memory capacity effects in the DRM paradigm. *Memory & Cognition*, *50*(7), 1443–1463. https://doi.org/10.3758/s13421-022-01298-y

Appendix

Criterion warning (from Miller et al., 2011)

You are now going to take another recognition test, but with different words. Again, some of the words will be words that were on one of the study lists, and some of the words will not have been on one of the study lists. As in the last test, write either "y" or "n" depending on if you recognise the word. However, this time I would like you to be very careful about saying "yes" to any word. Twelve out of the 14 lists that you heard were composed of words that formed a central theme. For example, the following list could have been presented: "mug, saucer, tea, measuring, coaster, lid, handle, coffee, straw, goblet, soup, stein, drink, plastic, sip." All of these words are related to "cup," but "cup" would not have been presented. Yet, very often subjects will falsely recognise these non-presented words like "cup," and they will be very confident that these words actually occurred. This task is meant to cause memory distortions, and it's meant to trip you up. In this test, I would like you to avoid saying "yes" and falsely recognising these non-presented words like "cup" as much as possible. A good rule to use in order to avoid falsely recognising a word is to be very careful in saying "yes" to any word that is strongly related to one of the study themes. If a particular word (like cup) fits one of the themes that you heard during the study session (cup-like words), then it is very likely that the word was not presented. If a word (like plastic) seems only weakly related or not related at all to one of the study themes, then it is likely that the word was presented. Again, I want you to be extremely careful about saying "yes" to words that are related to one of the study themes, because most of these words will not have been on one of the study lists that you heard.

Strong critical lure warning (from Neuschatz et al., 2001)

You should be cautious when taking this test. Our purpose in this experiment is to try to trick you into selecting items that weren't actually presented. To do that, we presented you with lists of thematically related words. For instance, you may have heard lists like: Writer, Poet, Novel, Book, etc. in which the word Author was never presented. Our purpose in presenting you with these lists was to try to get you to select the word Author even though it was never stated. Do your best to avoid being tricked in this way. One way to avoid being tricked is to carefully consider what characteristics you think you remember about the words. Much research has shown that presented and non-presented words can be reliably distinguished based on their characteristics. Here are some things to keep in mind:

1. Being confident, by itself, does NOT guarantee that you really heard the word. People are often quite confident that they remember items that were, in fact, never presented.
2. Remembering perceptual details (like the sound of the speaker's voice) makes it more likely that you really heard the word. You should be better able to recall details about the actual sound of the speaker's voice if the word is one that you really heard rather than one you only think you heard.
3. Remembering emotional details makes it more likely that you really heard the word. Sometimes words can produce an emotional response (make you feel happy, sad, angry, etc.). On average, your memory for words that were really presented should be more emotionally vivid than words you were not presented with.
4. Remembering contextual details makes it more likely that you really heard the word. For instance, if you can remember where in the list the item was presented (early, middle, late) it is likely to be a word that was really presented.