Cognition 279 (2027) 106782



Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit



Full Length Article

Probing spatiotemporal context reveals differences between true and false memories

Brynn E. Sherman *

Department of Psychology, The Ohio State University, United States of America

ARTICLE INFO

Keywords:
False memory
Temporal memory
Spatial memory
Spatiotemporal context
Deese-Roediger-McDermott paradigm

ABSTRACT

Our memories do not faithfully represent the past, but rather are reconstructive, rendering them susceptible to the intrusion of false memories. However, the nature of the false memories is debated: Do false memories reflect the retrieval of a genuine, erroneous memory trace, or are they the product of semantic associations? This question reflects the fact that although some findings suggest that false memories can be phenomenologically compelling and rich in details, others suggest that there are subtle cues that distinguish true and false memories. Here, we take a novel approach to this question by assessing memory for the spatiotemporal context of false memories — the *when* and *where* of those memories. We demonstrate that context memory can provide a probabilistic cue to whether a memory is true or false in the Deese-Roediger-McDermott paradigm. Across participants, the reliability of reported context memory (i.e., the consistency with which people report remembering when or where a false memory occurred) revealed differences between true and false memories. Additionally, confidence in memory for context (both within and across participants) differed between true and false memories — even when controlling for confidence in having encountered the false memory. However, these effects were malleable: Under certain circumstances, false memories appeared to have genuine context traces. Together, these data provide new insights into the circumstances under which true and false memories differ in subtle but meaningful ways, speaking to theories of how false memories may arise from the reconstructive nature of memory.

Our memories are not a veridical reflection of our past, but rather are a *reconstruction* of our experiences (Bartlett, 1932). The reconstructive nature of memory renders memory fallible in many ways (Schacter, 1999). Most obviously, we do not remember everything we experience: That is, our memories are prone to forgetting. However, beyond simply forgetting, our memories can be inaccurate in a variety of ways. For example, memories are susceptible to more subtle errors, or distortions (Schacter et al., 2011): We can erroneously recombine information across memories, or misremember certain details of an experience. And even beyond these more subtle distortions, our minds are also susceptible to the intrusion of *false memories*. That is, people can report having experienced events that never actually occurred (Johnson & Raye, 1998; Loftus, 2005; Pickrell et al., 2016; Roediger & Gallo, 2022). These false memories are not just induced within individuals (i.e., a false autobiographical memory; Loftus & Pickrell, 1995), but also can be reliably observed at the population level (with collective misrememberings of certain events; Maswood & Rajaram, 2019; Prasad & Bainbridge, 2022).

Experimentally, robust false memory formation has been studied most often using the Deese-Roediger-McDermott (DRM) paradigm, in which participants are presented with a list of semantically related words centered on an unpresented theme word (Deese, 1959; Roediger & McDermott, 1995). For example, a participant may be presented with a list of words centered around the theme of *sleep* (e.g., *bed, dream, nap,* etc.) but never actually be presented with the word *sleep*. When tested on their memory for the words, people tend to erroneously recall or falsely recognize the *critical lure* theme word which is related to the word list but was never actually presented (i.e., *sleep*). False recognition and recall rates to these critical lure words are high (often comparable to that of genuinely presented items), suggesting that this paradigm effectively induces false memories.

How genuine, though, are these false memories (Gallo, 2010; Jou & Flores, 2013)? Many have argued that false memories are not merely the result of increased familiarity (given the semantic relatedness of the list), but that these memories are qualitatively and experientially equivalent to true memories: Remembering a false memory is as vivid

* Corresponding author at: Psychology Building, Room 207, 1835 Neil Avenue, Columbus, OH 43210, United States of America.
E-mail address: sherman.1787@osu.edu.

https://doi.org/10.1016/j.cognition.2026.106782
Received 18 May 2026; Received in revised form 30 September 2026; Accepted 3 November 2026
Available online 7 November 2026
0010-0277/© 2026 The Author. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

and phenomenologically compelling as recalling a true memory (Lampinen et al., 1997). Evidence for such an interpretation comes from studies which have measured the subjective aspects of recollecting false memories. For example, critical lures are often recognized with high rates of subjective confidence (Roediger & McDermott, 1995), which cannot merely be attributed to a high degree of familiarity (i.e., the sense of "knowing" that a word has been seen, without a subjective sense of recollection; Payne et al., 1996; Norman & Schacter, 1997; see also Holmes et al., 1998). Further, participants report vivid and strong recollections of particular aspects of an experience. For example, people will report remembering specific details of the experience such as the source of the memory (i.e., the identity of the speaker of the critical lure word; Payne et al., 1996; Lampinen et al., 1999) or contextual features, such as when the critical lure was presented (Read, 1996; see also Norman & Schacter, 1997).

However, others have argued that there may be some signals which dissociate true and false memories (see Jou & Flores, 2013). For example, some studies have found that people have longer response times and lower confidence ratings in their memory for critical lures, compared to true memories (Jou, 2008; Jou, 2011; Jou et al., 2004). Additionally, although people do often report vivid subjective experiences associated with false memories, critical lures have been found to be rated as subtly, but reliably, less vivid (Mather et al., 1997; Neuschatz et al., 2001; Norman & Schacter, 1997). Together, these data suggest that there may be some signals within the memories themselves that allow people to implicitly detect whether a memory is true or false.

What might explain these discrepant findings? One possibility is that there are two distinct kinds of false memory arising from dissociable mnemonic processes (Parks et al., 2026). Specifically, some false memories may reflect *false familiarity* (i.e., an increased sense of familiarity, perhaps driven by the semantic associations among items), which could give rise to lower confidence in false memories. Other false memories may instead reflect *false recollection* (i.e., a genuine recollective process, driven by the erroneous binding of genuine episodic details associated with a true memory to the false memory), which could give rise to phenomenologically compelling false memories. Another notable distinction is that findings of phenomenologically powerful false memories primarily come from studies which directly ask participants to report those details (i.e., report the source of their memories; Read, 1996; Payne et al., 1996; Lampinen et al., 1999). In contrast, studies which find evidence for subtle dissociations typically ask participants to report aspects of *meta-memory* (i.e., how confident they are or how vivid their memories are; Jou et al., 2004; Norman & Schacter, 1997; Mather et al., 1997).

Here, we develop a novel approach for assessing whether true and false memories are phenomenologically distinct by more directly probing both memory and meta-memory for contextual features of an experience using the DRM paradigm. Specifically, we ask whether variability in memory for context (i.e., *when* and *where* something occurred, the hallmarks of 'episodic memory'; Tulving, 1993) — both across and within individuals — can provide subtle, but informative cues as to whether a memory is true or false. One possibility is that false memories will be devoid of contextual features, such that probing memory for context will reveal differences between true and false memories; this pattern of results would be consistent with the notion of false *familiarity* and would align with prior work demonstrating that there are some subtle differences between true and false memories (e.g., Jou & Flores, 2013). However, an alternative possibility is that false memories will indeed contain (false) contextual signatures; this pattern would be consistent with the notion of false *recollection* and would align with prior work demonstrating that false memories can indeed contain phenomenologically vivid detail, including for some contextual features (e.g., Lampinen et al., 1997; Read, 1996).

In an initial experiment aimed at teasing apart these possibilities, we presented participants with a list of words (modeled after the DRM paradigm) which were shown in a fixed order across participants. When testing participants' temporal memory for each word, we found that there was reliably greater variability in the timeline placements of critical lures, relative to true memories, indicating that — at the group level — false memories can be detected by examining the reliability in people's memories for temporal context. In a second experiment, we extended this finding using a similar paradigm in which one half of the list was associated with one theme and the second half of the list was associated with a second theme. Although we still observed numerically increased variability for critical lures, we also found that participants placed the critical lures near in time to their related words, suggesting that people can confabulate contextual features based on shared semantic context. In a third experiment, we sought to understand whether context memory can provide a cue to true vs. false memories not just at the group level (when aggregating across participants) but also within an individual. To this end, we had participants judge their subjective confidence in their temporal memory judgments, and we found that participants were reliably less confident in their temporal memory placements for critical lures (even when controlling for confidence in having seen the word). In a fourth experiment, we extended this by testing participants' memory for the spatial locations of words, rather than time, and similarly found that memory (and confidence) for spatial context differs between true and false memories. In a final experiment, we tested participants on their memory for both spatial and temporal context and found that critical lures were associated with decreased confidence in both dimensions. Together, these data suggest that both memory for spatiotemporal context (i.e., when and where an event occurred) itself, as well as confidence in context memory, can provide insight into whether memories are genuine.

1. Experiment 1: Variability in memory for 'when'

In a first experiment, we addressed whether variability in memory for temporal context (i.e., *when* in time something occurred) would dissociate true and false memories. We presented participants with a list of words and subsequently asked them to judge not only whether they saw each word, but also when in time they saw it (responding on a continuous timeline corresponding to the duration of the list). Critically, all words were presented in a fixed temporal order across participants, enabling us to examine the variability in temporal context memory (i.e., memory for *when*) across individuals. Given that critical lure words were not presented and thus do not have an associated temporal context to be remembered, we hypothesized that there would be greatest variability in memory for critical lures, compared to genuine old items. However, it is also possible that there could be a systematic bias (e.g., to remember the critical lures as having occurred early in the sequence; Read, 1996), which may predict no difference in variability for critical lures vs. old items.

1.1. Methods

1.1.1. Participants

Participants in all experiments (except Experiment 5b) were recruited via the Prolific platform. All participants (in all experiments) were adults 18–35 years old residing in the United States who were fluent in English. Additionally, all participants had an approval rate of at least 99% and had completed at least 500 prior Prolific submissions. All participants provided informed consent, and the study was approved by the relevant Institutional Review Board. Participants were paid at a rate of $15/h.

The final sample size was 50 participants after exclusions and replacement. Participants were excluded if (i) they did not false alarm to at least one of the two critical lures; (ii) they false alarmed to more than 50% of the unrelated foils; or (iii) their temporal memory judgments were poorly correlated with the true temporal position of the items (Spearman rank correlation <0.3). Due to an error with the data writing, one participant was missing data from three trials of the memory test.

We employed these (relatively strict) exclusion criteria to ensure that participants were doing the task in good faith; this was particularly important given the online data collection platform and the fact that there was no encoding task (or other attention checks) to serve as an independent metric of participants' engagement in the task.

To achieve the usable sample of 50, 158 participants were collected. Of the 158, 25 did not false alarm to at least one of the two critical lures (criterion i); 24 false alarmed to more than 50% of unrelated foils (criterion ii); and 88 exhibited Spearman rank correlations <0.3 (criterion iii). (One participant failed on criteria i & ii; 9 participants failed on criteria i & iii; 20 participants failed on criteria ii & iii). Additionally, one participant was excluded for completing the task twice. Although this exclusion rate is quite high, we again note that we purposefully employed strict criteria given the online nature of data collection. To assess the extent to which our results were robust to the specific exclusion criteria, we ran exploratory analyses (for all experiments) both without any exclusions, as well as with a more liberal Spearman rank correlation threshold (*rho* < 0). These analyses (for all experiments) are reported in the Supplemental Text.

To ensure that the sample size of $N = 50$ was sufficiently powered and appropriate for use in subsequent experiments, we performed a post-hoc simulation-based power analysis to quantify our ability to detect our primary effect of interest (a difference in the standard deviation in timeline placement for old items vs. critical lures). This analysis suggested that our achieved power was 87%.

1.1.2. Task design & procedure

The experiment was administered online via a web-based interface using custom JavaScript code. The task consisted of an encoding phase followed by a memory phase. Participants were informed that they would be viewing a sequence of words, followed by a memory test, though they were not informed of the nature of the memory test (i.e., they did not know that their temporal memory would be tested).

During the encoding phase, participants passively viewed a list of words, which were presented one at a time orthographically on the screen (Fig. 1A, left). Each word was presented on screen for 2 s, separated by a 1 s blank inter-stimulus interval. The word list consisted of 40 words total. The list included two intermixed word sets (15 items each; related to *window* and *smell*) that have been used in prior DRM studies; these word lists were specifically chosen as they are among the top four word lists associated with a high probability of false recall (Stadler et al., 1999; Pardilla-Delgado & Payne, 2017; Fig. 1A, right). Additionally, the first and last five words of the list were unrelated to the two lists, but were other words drawn from distinct DRM word sets. These items were placed at the beginning and end of the list (and were not tested during the memory phase) to reduce the influence of primacy and recency effects on memory for the words of interest.

Critically, the word list order was fixed across participants, such that all participants saw the same words in the same order. We implemented this fixed order (rather than randomizing the word order for each participant) so that we could examine the reliability in participants' temporal memory judgments for each word.

Following the encoding phase, participants completed a memory test in which they were tested on their memory for the words (Fig. 1B). For each word, participants were first asked to indicate whether they remember seeing the word (responding with "Yes" and "No" buttons on

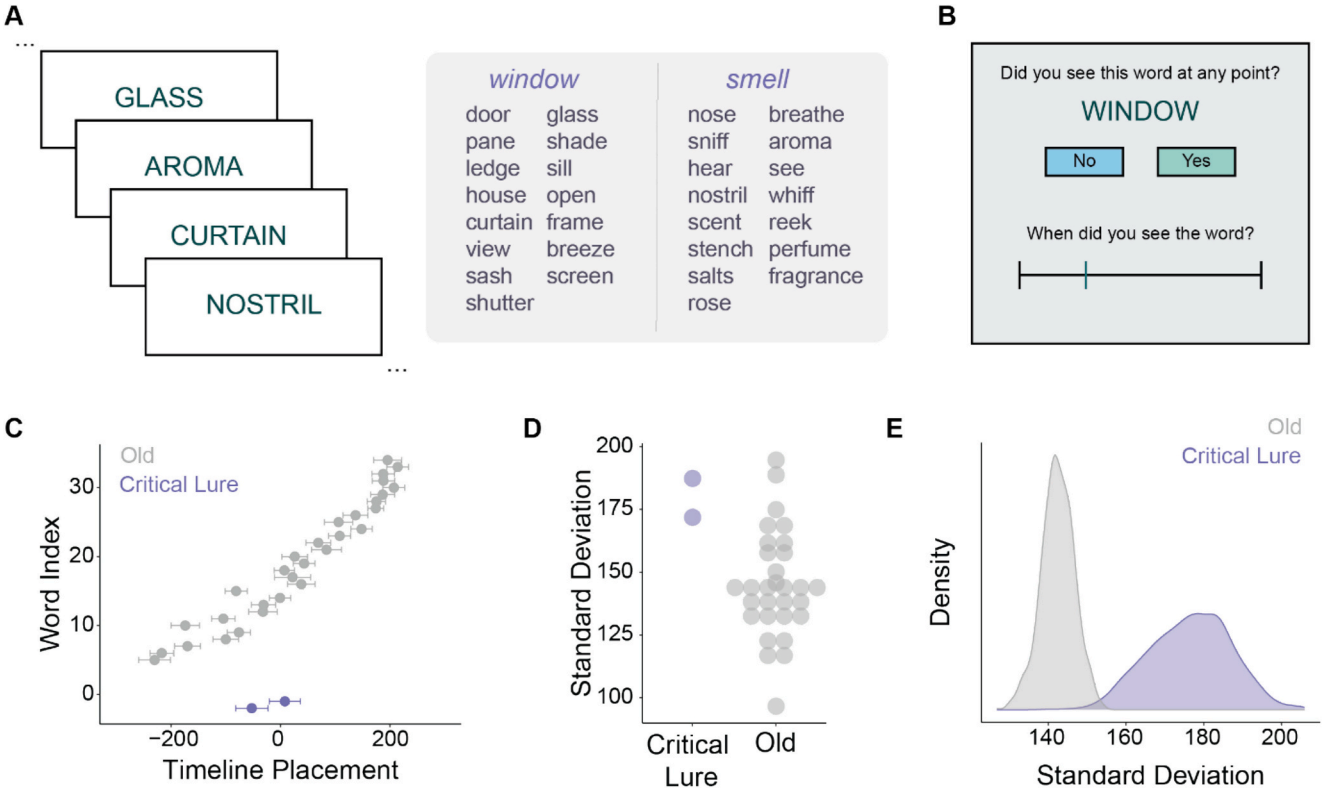


Fig. 1. *Experiment 1 Design and Data.* A) During encoding, participants viewed a list of words presented on the screen, one at a time (left). The words were drawn from two sets used in prior studies of the DRM paradigm (right; critical lure theme words shown at the top of each list in purple). B) At test, participants were asked to indicate whether they remembered seeing a given word, followed by a temporal context memory judgment in which they moved a marker along a timeline to indicate when they remember seeing the word. C) Average timeline placements (along the x axis) for each old word (grey), ordered according to the actual list position at encoding (along the y axis), as well as the two critical lure words (purple), which are indicated as Index = −1. Error bars indicate the standard error of the mean across participants. D) Standard deviation in timeline placement across participants for each critical lure (purple) and old word (grey). E) Population distribution estimates of the mean standard deviation for old and critical lure words, obtained via bootstrap resampling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the screen). Then, regardless of whether they indicated remembering seeing the word, participants were asked to indicate when in time they recall seeing the word. Participants responded on a timeline which was bound by the labels "Beginning of the task" on the lefthand side and "End of the task" on the righthand side. The timeline was 780 pixels wide, and responses were coded as ranging from −390 to 390 along the x-axis. Participants could click and drag a response marker along the timeline, and they pressed the space bar to register their response. The memory test was self-paced.

Prior to the memory phase, participants were instructed: "In this next part of the task, you're going to be asked about those words. For each word, you'll be asked whether you saw it before or not. Regardless of whether you answer yes or no, you'll then be asked to indicate **when** you saw it. A timeline will be visible on the screen, and you can click on the timeline and then press spacebar to submit your response. That's all there is to it. The task might be challenging, but just do your best." On each memory trial, participants were prompted with the question "When did you see the word? (Drag and drop on the timeline, then press spacebar to submit your answer)".

Participants were tested on their memory for all words in the two presented DRM sets (15 words each; 30 "old" words total), along with the two critical lures (one for each set; *window* and *smell*). We additionally included 3 unrelated foil words which were not associated with the two sets but were drawn from separate DRM word sets. All words at test were presented in a random order for each participant.

1.2. Results

We first assessed participants' recognition memory for the words. On average, participants correctly identified 85.0% of words as old (SD = 12.2%). False alarm rate to unrelated foils was low (7.33%; SD = 13.9%). The critical lure words were falsely recognized as old 77.0% of the time (SD = 25.2%). Reflecting these patterns, participants were significantly more likely to falsely recognize critical lures, compared to unrelated foils, $t(49) = 17.92$, $p < 0.001$, $d = 2.53$. There was also a small, but significant difference in recognition rates for critical lures, relative to old items, $t(49) = 2.21$, $p = 0.032$, $d = 0.31$, such that critical lures were indeed endorsed as 'old' at a lower rate than genuinely old items. Nevertheless, these patterns verify that the paradigm was effective in inducing false memories.

Our primary interest was in examining participants' temporal memory judgments. We first computed the Spearman rank correlation between participants' responses along the timeline and the actual temporal positions of the old items. On average, participants exhibited a correlation of *rho* = 0.54 (SD = 0.16), indicating that they were roughly able to reconstruct the temporal order of the sequence (Fig. 1C).

We next assessed the reliability of timeline responses across participants by computing the standard deviation in timeline placement for each word. We were specifically interested in whether the reliability of timeline responses differed for old words vs. critical lures (we did not consider unrelated foils in this analysis). We conducted the analysis specifically on trials in which a word was endorsed as 'old' to ensure that any differences in variability could not be attributed to lack of recognition memory. On average, the standard deviation in timeline placement across participants for old words was 145.2 pixels and for critical lure words was 179.6 pixels (Fig. 1D). This difference cannot be explained by large differences in recognition rates for critical lures vs. old items: There was an average of 38.5 responses going into the standard deviation computations for critical lures and an average of 42.4 going into the standard deviation computations for old items.

To assess the statistical reliability of this difference, we ran a bootstrap resampling test. Across 1000 iterations, we randomly resampled the 50 participants with replacement and recomputed the standard deviations for critical lure and old words. We then derived an empirical *p* value by assessing the proportion of iterations for which the average standard deviation for the old words was lower than the average standard deviation for critical lure words. We indeed found a significant difference (bootstrap $p = 0.002$; Fig. 1E), such that participants' timeline responses to critical lure words were significantly more variable than their responses to the old items, consistent with the fact that there is no genuine temporal context signal for critical lure words.

2. Experiment 2: manipulating variability in temporal memory

The results of Experiment 1 suggest that — at the group level — variability in mnemonic judgments of temporal context (i.e., by asking *when* a given memory occurred) provides a cue to whether a memory is true or false. These results perhaps are not surprising: Given that there is no true temporal context for critical lures, then it almost necessarily should be the case that there should be greater variability in participants' memories for those items, assuming that participants are able to recall the temporal context of actually presented items. However, this finding does perhaps stand in contrast to prior findings of a systematic bias to remember critical lures as having occurred at the beginning of the list (Read, 1996).

In Experiment 2, we sought to push around this effect, asking whether we could decrease the variability in the remembered temporal positions of critical lures. We adopted the same experimental design as Experiment 1, except that we organized the word list such that all words from one set were presented in the first half of the list and all words from another set were presented in the second half of the list. We reasoned that participants may use the mappings of the semantic context onto temporal context as a cue (see, e.g. work on 'content borrowing'; Lampinen et al., 2005; Arndt, 2006; Nieznański et al., 2018), guiding more precise estimates of temporal positions, even for critical lures.

2.1. Methods

2.1.1. Participants

A new sample of 50 usable participants was included in Experiment 2 (applying the same exclusion criteria as Experiment 1). Due to an error with data writing, three participants were missing one, two, and three trials respectively.

To achieve the usable sample of $N = 50$, we collected data from 80 participants. Four participants did not false alarm to at least one of the two critical lures; 9 had an unrelated foil false alarm rate > 0.5, and 23 had Spearman rank correlation <0.3 (1 participant exhibited both correlation <0.3 and no false alarms to the critical lures; 5 participants exhibited both correlation <0.3 and unrelated foil false alarm >0.5).

2.1.2. Task design & procedure

The experiment was identical to Experiment 1, except that the two word sets of interest were presented in a blocked manner. That is, rather than intermixing the two sets, participants first saw all words associated with *window* (in the first half of the list) and then all words associated with *smell* in the second half of the list (Fig. 2A). As in Experiment 1, there were five unrelated words at the start and end of the list to reduce the impact of primacy and recency effects.

2.2. Results

On average, participants correctly identified 84.6% of words as old (SD = 10.9%). False alarm rates to unrelated foils were low (6.0%; SD = 12.9%). The critical lure words were falsely recognized as old 90.0% of the time (SD = 20.2%). Again, the false alarm rate for critical lures was significantly higher than that of unrelated foils, $t(49) = 23.17$, $p < 0.001$, $d = 3.28$. In contrast to Experiment 1, we did not find a statistical difference in recognition rates for actual old vs. critical lures (if anything, it trended in the opposite direction, with numerically higher endorsement of critical lures as old, relative to actually old items; $t(49) = 1.87$, $p = 0.068$, $d = 0.26$), again verifying that the paradigm was effective in inducing false memories. Additionally, participants were

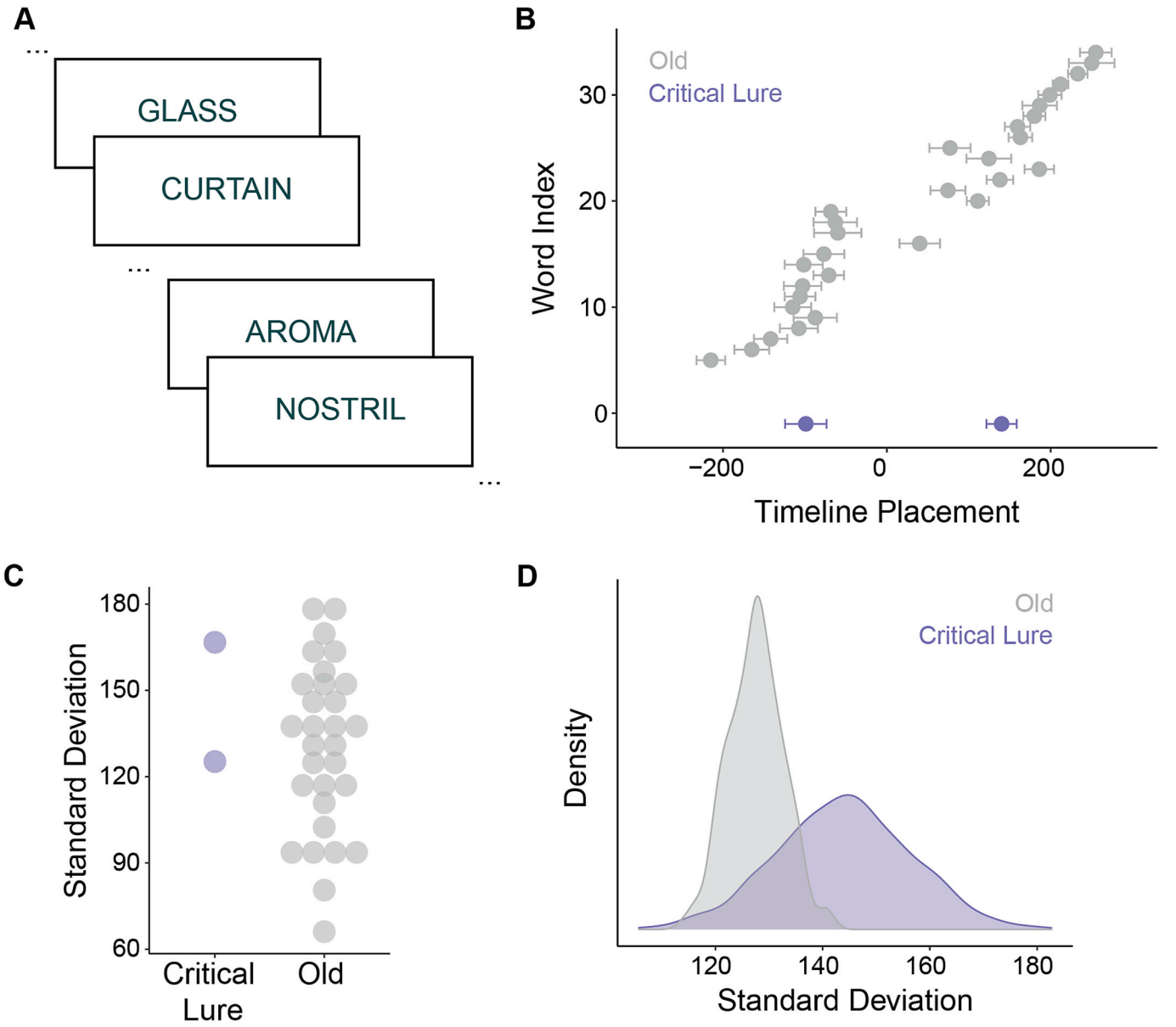


Fig. 2. *Experiment 2 Design and Data.* A) The procedure was identical to that of Experiment 1, except that during the encoding phase, participants viewed the word lists in a blocked order (all words associated with *window*, followed by all words associated with *smell*). B) Average timeline placements for each old word (grey), ordered according to the actual list position at encoding, as well as the two critical lure words (purple), which are indicated as Index = −1. Error bars indicate the standard error of the mean across participants. C) Standard deviation in timeline placement for each critical lure (purple) and old word (grey). D) Population distribution estimates of the mean standard deviation for old and critical lure words, obtained via bootstrap resampling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

able to reconstruct the relative temporal order of the old items, with an average *rho* = 0.62 (SD = 0.19; Fig. 2B).

When assessing the reliability of timeline placements across participants, we numerically replicated the pattern observed in Experiment 1, with higher variability (indexed by standard deviation across participants; Fig. 2C) in temporal memory for critical lures (SD = 146.0 pixels), compared to old items (SD = 130.1 pixels), though the difference was not significant (bootstrap $p = 0.071$; Fig. 2D).

That said, this weakened effect was largely expected, as we hypothesized that blocking the word list by topic set could reduce the variability in participants' temporal memory judgments (as they could leverage the mapping of the semantic context onto time to infer the approximate time of the critical lures). To quantify this, we asked whether there was a reliable difference in participants' timeline placements for the two critical lure words. Indeed, for participants who indicated remembering both critical lure words ($N = 39$), those participants reliably placed *window* as having occurred earlier in time than *smell*, $t(38) = 7.80$, $p < 0.001$, $d = 1.25$, consistent with the structure of the task (*window*-related words were presented in the first half of the list, and *smell*-related words were presented in the second half of the list). In contrast, there was no such difference in Experiment 1 (when the two sets were intermixed; $t(26) = 0.11$, $p = 0.92$, $d = 0.021$). Considering the two experiments together, there was a significant experiment x word interaction, $F(1, 64) = 21.2$, $p < 0.001$, verifying that the effects indeed differed across the two experiments. Together, these data suggest that variability in temporal memory for false memories is malleable: Under certain circumstances, people may reliably confabulate when they experienced an event.

3. Experiment 3: measuring confidence in memory for 'when'

Experiments 1–2 suggest that variability in temporal context memory at the group level can be used to draw insight into whether a memory is genuine or false. In Experiment 3, we sought to extend this by asking whether there are also *within*-participant signals of context memory that differentiate these distinct types of memory. To this end, we adopted a similar approach to Experiments 1–2 but additionally asked participants to rate their subjective confidence in their timeline placements for each word.

3.1. Methods

3.1.1. Participants

A new sample of 50 usable participants was included in Experiment 3. Participants were excluded if (i) they did not false alarm to at least one of the four critical lures; (ii) their average memory score (see below) to unrelated foils was >0; or (iii) their temporal memory judgments were poorly correlated with the true temporal position of the items (Spearman rank correlation <0.3).

Data were collected from 123 participants in order to achieve a usable sample size of $N = 50$. 8 participants did not false alarm to at least one critical lure; 6 participants had an average memory score > 0 for unrelated foils; and 65 participants exhibited a Spearman rank correlation <0.3. Six participants exhibited both foil memory score > 0 and rank correlation <0.3.

3.1.2. Task design & procedure

The experimental design was similar to that of Experiment 1, but with several critical changes to optimize the task for probing confidence in temporal memory.

First, rather than viewing one word list (composed of two DRM sets, as in Experiments 1–2), followed by a memory test, participants completed four study-test rounds, such that they viewed one word list (containing one DRM set of 15 words; *window*, *smell*, *sleep*, and *doctor*) and then were immediately tested on their memory for that list, before moving onto the next block. We included 3 unrelated words at the beginning and ending of each list (to attenuate primacy/recency effects), such that each list contained 21 words in total. As in Experiments 1 and 2, all words were presented in a fixed order across participants. Each test phase was specific to its preceding study block, such that only words from that given block were tested, and the 'left hand' end of the timeline would correspond to the beginning of *that block* (rather than the beginning of the entire experiment).

Second, because participants completed four study-test rounds, the temporal memory test could no longer be a surprise (as it was in Experiments 1–2). We thus told participants in advance of the first list that their memory would be tested not only for whether they saw the words but also for when in time they saw them.

Third, we reduced the encoding time of each word to 1500 msec. This slight decrease in encoding time, relative to the 2000 msec in Experiments 1–2, was meant to account for the fact that (i) participants were viewing fewer words per list, and thus it may be easier for them to

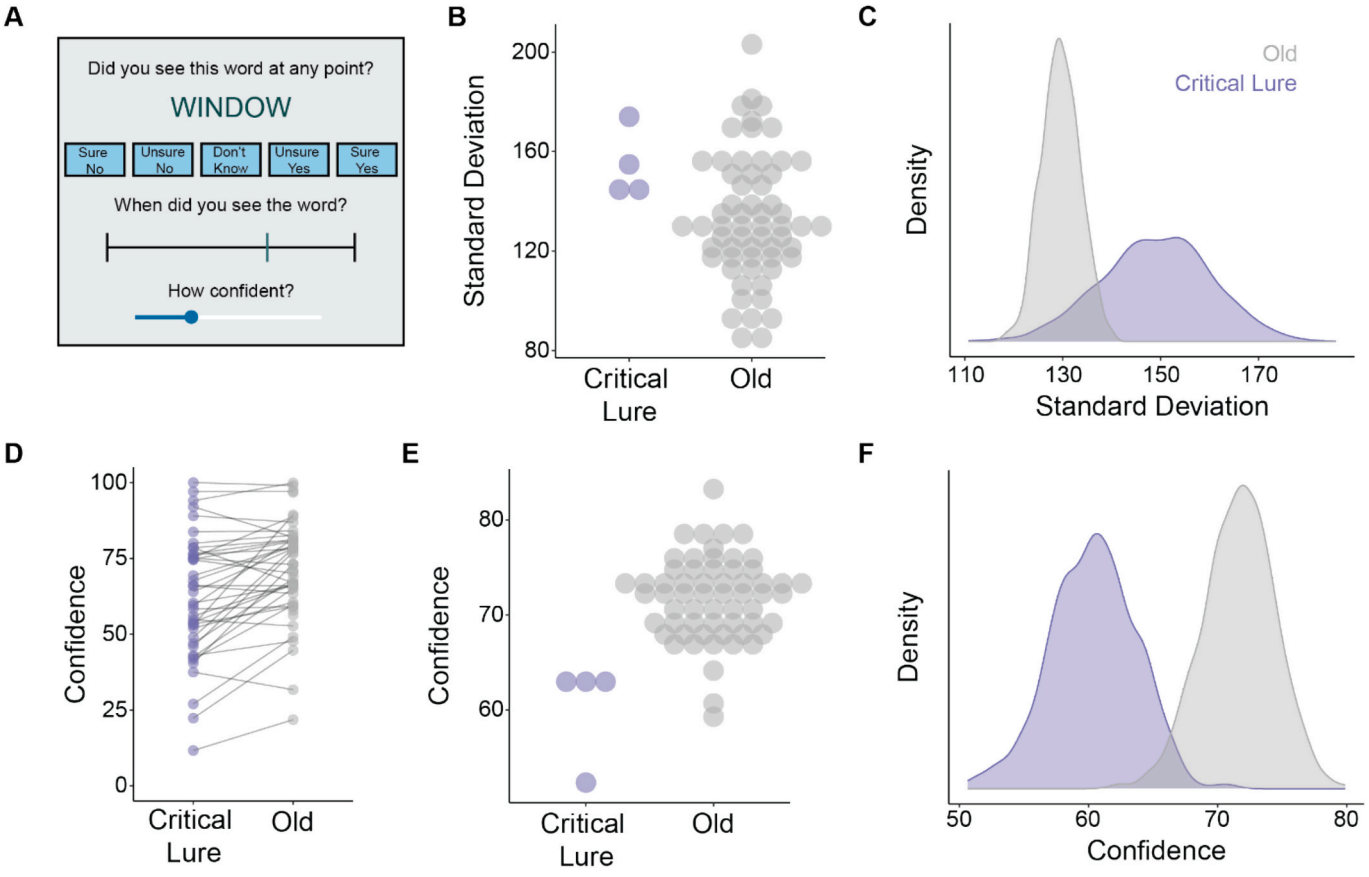


Fig. 3. *Experiment 3 Design and Data.* A) During the memory test, participants made confidence judgments both for whether they recognized the word (using a 5-point response scale), as well as for their timeline placements (using a slider bar). B) Standard deviation in timeline placement for each critical lure (purple) and old word (grey). C) Population distribution estimates of the mean standard deviation for old and critical lure words, obtained via bootstrap resampling. D) Mean timeline confidence ratings for critical lures vs. old items, separately for each participant (each dot/line indicates an individual participant). E) Mean confidence ratings for each word (separately for critical lures, purple, and old words, grey). F) Population distribution estimates of the mean confidence ratings for old and critical lure words, obtained via bootstrap resampling. Note that the data plotted in B—F are all plotting "Sure Yes" recognition responses only. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

remember each word with less encoding time; and (ii) the overall task duration was longer, given that participants encountered more total words across the four study-test rounds.

Fourth, during the memory test (Fig. 3A), we gave participants five response options when asking them whether they remember seeing the word. Rather than simply responding "yes" or "no" as in Experiments 1–2, participants were given buttons with the options "Sure No", "Unsure No", "Don't Know", "Unsure Yes", and "Sure Yes". The purpose of this change was to measure confidence in the word recognition itself so that we could assess whether any differences in temporal memory confidence could be attributed to differences in recognition confidence. Participants were encouraged to respond specifically about their confidence in seeing the word, regardless of whether they remembered when the word occurred.

Fifth, following each temporal memory judgment, participants were asked to rate their confidence in their temporal memory. They were presented with a sliding scale which ranged from 0 to 100. They could click and drag along the slider and pressed the space bar to submit their response. As in Experiments 1–2, the memory test was self-paced.

3.2. Results

During the memory test, participants first indicated their recognition memory for the word by responding "Sure No", "Unsure No", "Don't Know", "Unsure Yes", or "Sure Yes". These responses were coded as −2, −1, 0, 1, and 2, respectively. On average, the mean memory score for old words was 1.51 (SD = 0.29), indicating that participants most often responded "Unsure Yes" or "Sure Yes" to old items. In contrast, the memory score for unrelated foils reflected a likelihood to respond "Unsure No" or "Sure No" (mean memory score = −1.41, SD = 0.42). The mean memory score for critical lure words was 1.00 (SD = 0.83). Participants were significantly more likely to endorse critical lures as old, compared to unrelated foils, $t(49) = 19.9$, $p < 0.001$, $d = 2.82$. Additionally, the mean memory score for old items was significantly greater than that of critical lures, $t(49) = 3.97$, $p < 0.001$, $d = 0.56$. Thus, although participants were likely to endorse critical lures as old, they were perhaps less confident in having seen those words, relative to old items, replicating past work (Jou et al., 2004).

To assess overall temporal memory, we again computed the Spearman rank correlation between participants' responses along the timeline and the actual temporal positions of the old items. On average, participants had a Spearman *rho* = 0.50 (SD = 0.19). When assessing the reliability of timeline placements across participants, we replicated the pattern observed in Experiment 1, with significantly higher variability (indexed by standard deviation across participants) in temporal memory for critical lures, compared to old items ($p = 0.007$). Importantly, this effect held even when limiting the analysis to words in which participants indicated the highest level of recognition confidence ("Sure Yes"; $p = 0.046$). Although the effect was numerically weaker, it is worth noting that fewer participants were included in this analysis (40, as opposed to 50) by virtue of limiting to "Sure Yes" responses (Fig. 3B-C).

Our critical interest in Experiment 3 was examining participants' confidence in their temporal memory judgments. When considering only the items that participants indicated seeing (both "Unsure Yes" and "Sure Yes" responses), we found that participants were significantly less confident in their temporal memory for critical lures, relative to old items, $t(49) = 5.72$, $p < 0.001$, $d = 0.81$. Importantly, this effect was not merely driven by the fact that participants were also more likely to indicate that they were unsure whether they saw the critical lures (given the overall greater proportion of "Unsure Yes" responses to critical lures). When limiting the analysis only to trials in which participants responded "Sure Yes", we observed the same pattern, $t(39) = 4.62$, $p < 0.001$, $d = 0.73$ (Fig. 3D). Thus, even when participants have high-confidence false recognition memories, their patterns of confidence in indicating *when* they saw the word provide a window into whether the memories are genuine.

Finally, we also assessed whether confidence in temporal context memory differentiates true and false memories at the group-level (similar to the timeline variability analysis above). Specifically, rather than computing the mean confidence rating across words for each participant, we computed the mean confidence rating for each word, across participants (Fig. 3E). We then asked whether there was a systematic difference in confidence ratings for critical lures vs. old items. Indeed, we found that critical lures had a lower mean confidence rating than old items, both when considering all 'yes' responses and when only considering "Sure Yes" responses (bootstrap $ps < 0.001$; Fig. 3F).

4. Experiment 4: assessing memory for 'where'

Thus far, we have examined memory for temporal context – that is *when* an event occurred in time. In Experiment 4, we extended this approach, probing memory for spatial context (i.e., *where* an event happened). To this end, we adapted the experimental design of Experiment 3, except that participants encoded words that varied in their spatial location on the screen, and then were asked to recall where in space they remembered seeing the word.

4.1. Methods

4.1.1. Participants

A new sample of 50 usable participants was included in Experiment 4. Participants were excluded if (i) they did not false alarm to at least one of the four critical lures; (ii) their average memory score to unrelated foils was >0; or (iii) their accuracy on the spatial memory judgments were at chance (proportion correct <0.25).

Data were collected from 60 participants in order to achieve a usable sample size of $N = 50$. Two participants did not false alarm to at least one critical lure; 3 participants had an average memory score > 0 for unrelated foils; and 6 participants exhibited spatial memory accuracy <0.25. One participant exhibited both a foil memory score > 0 and spatial memory accuracy <0.25.

4.1.2. Task design & procedure

The experimental design was similar in structure to that of Experiment 3, in that participants completed four study-test rounds, with each round associated with a different DRM set (we used the same sets and the same list orders as in Experiment 3). However, instead of testing participants on their temporal memory, we tested participants on their memory for the spatial context.

During encoding, the words were presented on a square, with each word appearing in the center of one of the four quadrants of the square (though note that there were no visual demarcations of the quadrants; Fig. 4A, top). As in Experiment 3, because participants completed four study-test rounds, they were told in advance of the first list that they would be tested on their memory both for whether they saw the word, as well as where it appeared on the screen. Each word appeared on the screen for 1500 msec, followed by a 1000 msec blank ISI, as in Experiment 3. All words were presented in a fixed presentation order and at fixed spatial positions across participants.

During the memory test, participants were again first asked to indicate whether they remember seeing the word, using the 5-point response scheme ("Sure No", "Unsure No", "Don't Know", "Unsure Yes", "Sure Yes"). Then, instead of being asked to indicate when they remember seeing the word, participants indicated where on the screen they think they saw the word. They were given a square, with the four quadrants demarcated, and they used the mouse to click in one of the four quadrants to register their response (Fig. 4A, bottom).

Following the spatial memory judgment, participants were asked to rate their confidence in their spatial memory judgment, using the same confidence slider interface as in Experiment 3.

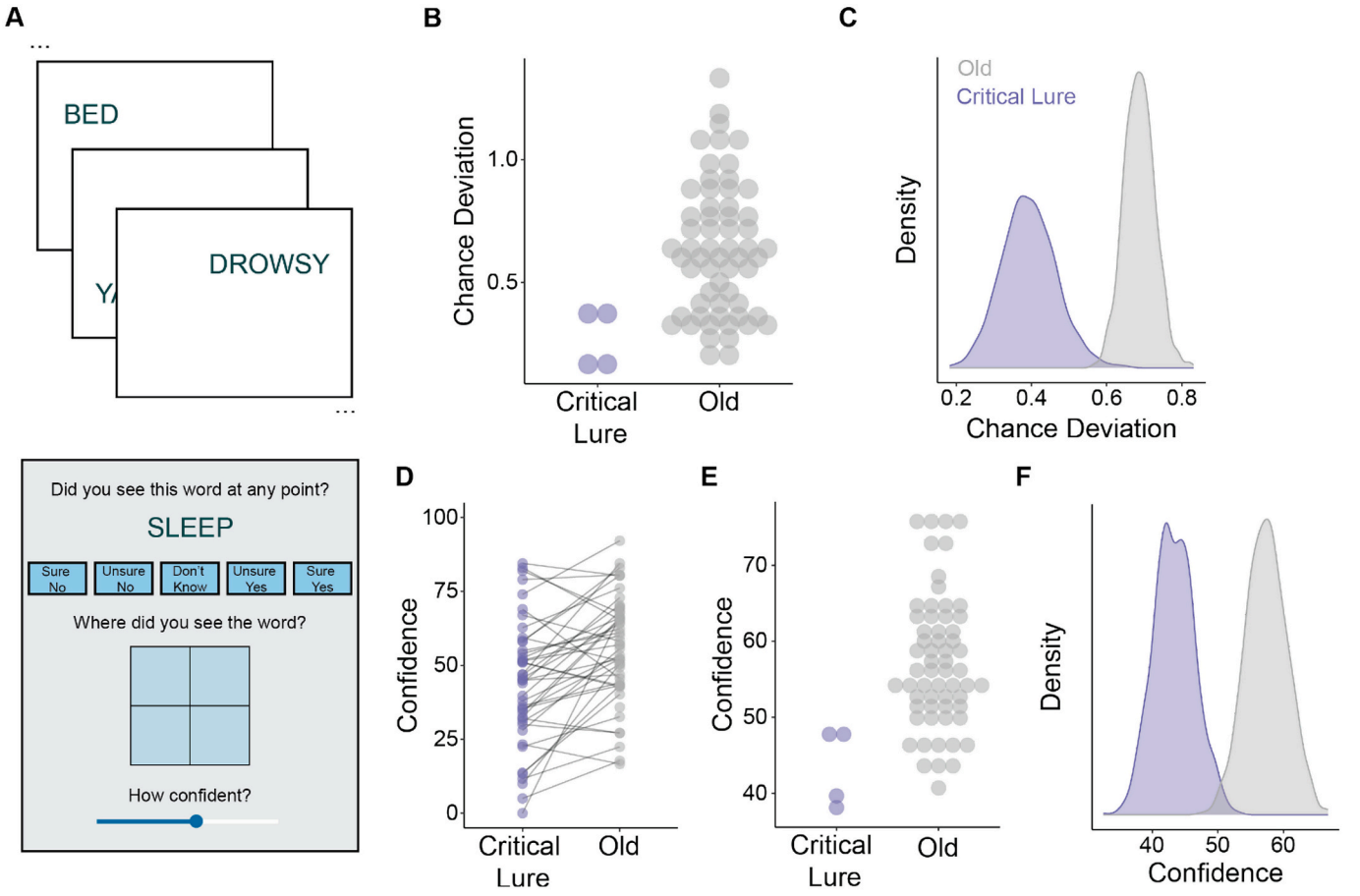


Fig. 4. *Experiment 4 Design and Data.* A) During encoding (top), participants viewed words on a square, with each word presented in the center of one of the four quadrants. During the memory test (bottom), participants indicated their spatial memory by clicking which quadrant they remember seeing the word, followed by a confidence rating. B) Spatial memory chance deviation (quantifying the relative consistency in quadrant responses across participants; higher numbers indicate greater consistency) for each critical lure (purple) and old word (grey). C) Population distribution estimates of the mean chance deviation for old and critical lure words, obtained via bootstrap resampling. D) Mean spatial confidence ratings for critical lures vs. old items, separately for each participant (each dot/line indicates an individual participant). E) Mean spatial confidence ratings for each word (separately for critical lures, purple, and old words, grey). F) Population distribution estimates of the mean confidence ratings for old and critical lure words, obtained via bootstrap resampling. Note that the data plotted in panels B—F are for "Sure Yes" recognition responses only. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Results

On average, the mean memory score for old words was 1.34 (SD = 0.28), whereas the memory score for unrelated foils was −1.50 (SD = 0.45), indicating that participants successfully discriminated old and new items. The mean memory score for critical lure words was 1.19 (SD = 0.69). Replicating Experiment 3, participants were significantly more likely to endorse critical lures as old, compared to unrelated foils, $t(49) = 28.72$, $p < 0.001$, $d = 4.06$. However, there was no statistical difference in memory score for critical lures vs. old items, $t(49) = 1.42$, $p = 0.16$, $d = 0.20$. Assessing overall performance on the spatial memory test, participants correctly identified the quadrant that a word appeared in 46.4% of the time (SD = 14.2%).

We next sought to assess the reliability of participants' spatial memory judgments for each word. Using a similar logic to the timeline variability analysis in Experiments 1–3, we hypothesized that participants would exhibit greater variability in their memory for the spatial locations of critical lure words (given that there is no true spatial context in which they were encountered). To quantify this for each word, we computed the proportion of participants that remembered the word as having occurred in each of the four quadrants. For each quadrant, we then computed a "chance deviation score" by taking the absolute value of the difference between the proportion of times the word was selected in that quadrant and 0.25 (chance); we then summed those deviation scores across the four quadrants for each word. If a given word was chosen as being in all four quadrants with equal likelihood across participants, then the total deviation score would be 0, reflecting genuine chance performance across participants; higher values thus indicate that the pattern of responses across the four quadrants exhibited a greater deviation from chance (i.e., more reliable responses across participants). We then examined whether there was a difference in deviation for critical lures vs. old items (note that because of the nature of this analysis, the expected pattern of results is that the critical lures would have *lower* values than old items, whereas in the timeline tasks, we expected higher standard deviation values for critical lures). As with the temporal memory analysis, we assessed the statistical reliability of this difference with a bootstrap resampling test (limiting the analysis only to trials in which participants indicated that they had seen the word; "Unsure Yes" and "Sure Yes" responses). Indeed, we found that there was significantly lower deviation for the critical lures than the old items (bootstrap $p < 0.001$), confirming that there was no genuine spatial memory signal for critical lures. This was observed also when limiting the analysis only to trials in which participants responded to "Sure Yes" to seeing the word during encoding (bootstrap $p < 0.001$; Fig. 4B-C). This analysis suggests that, just as variability in temporal context judgments across participants differed between true and false memories, variability in spatial context judgments across participants can also tease apart true and false memories.

Again, we were also interested in assessing participants' confidence in their spatial memory judgments, and we replicated the patterns observed with temporal memory confidence in Experiment 3. When considering only the items that participants indicated seeing ("Unsure Yes" and "Sure Yes" responses), we found that participants were significantly less confident in their spatial memory for critical lures, relative to old items, $t(49) = 5.43$, $p < 0.001$, $d = 0.77$. This effect held when limiting the analysis only to trials in which participants responded "Sure Yes", $t(42) = 4.47$, $p < 0.001$, $d = 0.68$ (Fig. 4D). Thus, we again observed a dissociation between confidence in seeing a word and confidence in memory for its context: Even when participants are (falsely) highly confident in having seen a word, their confidence in their memory for contextual features distinguishes that memory from genuine ones.

Taking a complementary approach, we again computed mean confidence for each word (averaged across participants). We again found that there was reliably lower confidence for critical lure words, compared to old words (considering both all 'yes' and 'sure yes' responses only; bootstrap $ps < 0.001$; Fig. 4E-F). Together, these data suggest that probing participants' memory for *where* an event occurred can also provide unique insight into whether a memory is genuine.

5. Experiment 5: testing memory for 'where' and 'when'

Finally, we tested memory for spatial and temporal context simultaneously. That is, we asked people to indicate both when and where they believed the critical lure occurred. The rationale here is that by combining these pieces of information, it may be possible to identify false memories with much greater fidelity. We assessed this both in an online sample, as in the prior experiments (Experiment 5a), as well as an in-person sample (Experiment 5b).

5.1. Experiment 5a – online sample

5.1.1. Methods

5.1.1.1. Participants. Twenty-one usable participants were included in Experiment 5a. We opted for a smaller sample size in this study given that the experiment was longer in duration and there were more exclusion criteria; this sample size was chosen based on a power analysis considering the effect sizes for the difference in confidence ratings in both Experiments 3 and 4 (power analysis conducted on estimated $d = 0.7$, with 85% power).

For Experiment 5a, participants were excluded if (i) they did not false alarm to at least one of the four critical lures; (ii) their average memory score to unrelated foils was >0; or (iii) their accuracy on the spatial memory judgments were at chance (proportion correct <0.25); or (iv) their temporal memory judgments were poorly correlated with the true temporal position of the items (Spearman rank correlation <0.3).

73 participants were collected to achieve the final sample size of $N = 21$. One participant did not false alarm to at least one critical lure; three participants had a memory score > 0 for unrelated foils; 7 exhibited spatial memory accuracy <0.25; and 50 exhibited a Spearman rank correlation <0.3. Two participants exhibited both foil memory score > 0 and spatial memory accuracy <0.25; five participants failed on both spatial memory accuracy and Spearman rank correlation; and one participant failed on foil memory score, spatial memory accuracy, and Spearman rank correlation.

5.1.1.2. Task design & procedure. The experimental design combined Experiments 3 and 4. The encoding phase was identical to that of Experiment 4. During the memory phase, participants were tested on their memory for both spatial and temporal context. After indicating whether they remembered seeing a word (same response options as Experiment 3–4), they first indicated the quadrant they remember seeing the word, followed by a confidence judgment for space. They then indicated their memory for the temporal position of the word, followed by a confidence judgment for time.

5.1.2. Results

The mean memory score for old words was 1.54 (SD = 0.32), whereas the memory score for unrelated foils was −1.46 (SD = 0.32). The mean memory score for critical lure words was 0.87 (SD = 1.02). Participants were significantly more likely to endorse critical lures as old, compared to unrelated foils, $t(20) = 10.28$, $p < 0.001$, $d = 2.25$. Further, participants exhibited a lower memory score for critical lures, compared to old items, $t(20) = 2.66$, $p = 0.015$, $d = 0.58$. Overall, participants exhibited an average accuracy of 54.8% on the spatial memory test (SD = 18.5%) and exhibited an average correlation between their timeline responses and the true temporal positions of $rho = 0.49$ (SD = 0.20).

We next examined the variability in spatial and temporal context judgments across participants. Considering the reliability in spatial context judgments, we replicated the pattern observed in Experiment 4, with greater deviation from chance for old, compared to critical lure items (bootstrap $p = 0.046$); however, this did not hold when limiting the analysis only to "Sure Yes" responses (bootstrap $p = 0.64$). We did not replicate the patterns of greater variability in temporal context memory judgments, either for all 'yes' responses (bootstrap $p = 0.77$) or for "Sure Yes" responses only (bootstrap $p = 0.96$). Although we did not fully replicate both expected patterns of results, it is worth noting that the sample size in these analyses was less than half of what we included in Experiments 1–4. These analyses are likely highly dependent on sample size, as including fewer participants will give inherently noisier estimates of the population standard deviations (and individual outlying values can have an outsize influence). That said, we opted for a smaller sample size in this study (to optimize for the confidence-based analysis), understanding that it may weaken our ability to detect effects with the variability measures.

Crucially, however, we replicated the results of the confidence analyses from Experiments 3 and 4, finding that participants were less confident in their judgments of both spatial and temporal context for critical lures items. Specifically, we found that participants were significantly less confident in their spatial memory judgments for critical lures, compared to old items, both for all 'yes' responses, $t(20) = 4.38$, $p < 0.001$, $d = 0.96$, and when limiting the analysis only to "Sure Yes" responses, $t(16) = 3.36$, $p = 0.0040$, $d = 0.81$ (Fig. 5A). This pattern was mirrored in temporal memory judgments as well: We observed lower confidence for critical lures, compared to old items (all 'yes' responses: $t(20) = 3.07$, $p = 0.0061$, $d = 0.67$; "Sure Yes" responses only: $t(16) = 2.56$, $p = 0.021$, $d = 0.62$; Fig. 5B). Further, when assessing the mean confidence rating for each word across participants, we found reliably lower confidence in both spatial and temporal memory judgments (all bootstrap $ps < 0.001$).

Finally, we assessed whether judgments of context for spatial and temporal context co-vary with one another, and whether decreased confidence across both dimensions may distinguish true and false memories. To this end, we computed – for each word (excluding unrelated foils) – the average confidence in spatial and temporal memory judgments across participants. Across all words (considering 'yes' responses only), we observed a strong correlation between spatial and temporal memory confidence, $r = 0.70$, $p < 0.001$. Critically, when visualizing the relationship between spatial and temporal confidence, it was apparent that the critical lures dissociate from the majority of the datapoints (i.e., they are clustered in the bottom left corner, reflecting low confidence in both space and temporal judgments). We quantified this bivariate difference (i.e., that critical lures were lower in *both* spatial and temporal confidence) in two complementary ways: First, we ran a one-way MANOVA, which revealed a significant effect of word type (critical lure vs. old) on the confidence metrics, $V = 0.36$, $F(2, 61) = 17.28$, $p < 0.001$. Additionally, we computed the Mahalanobis

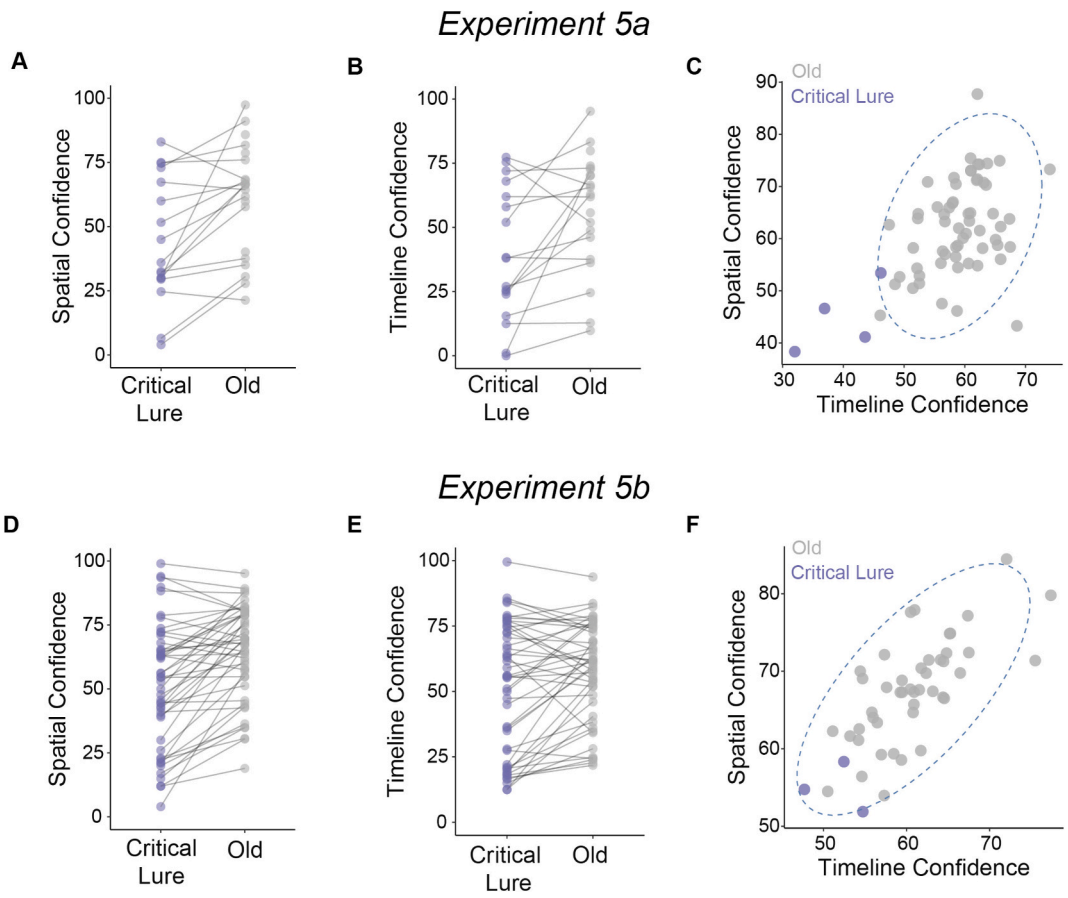


Fig. 5. *Experiment 5a (panels A—C) and Experiment 5b (panels D—F) Data.* A & D) Mean spatial confidence ratings for critical lures vs. old items, separately for each participant (each dot/line indicates an individual participant). B & E) Mean timeline confidence ratings for critical lures vs. old items, separately for each participant (each dot/line indicates an individual participant). C & F) Mean confidence rating for both spatial and temporal judgments (averaged across participants) for each word. Dotted circle indicates the 95% boundary of the 'old' distribution, as computed via Mahalanobis distance. All plotted data are for "Sure Yes" recognition responses only. In panel (F), the words from block 1 are not included in the plot.

distance (quantifying the multi-dimensional distance) for all words, relative to the centroid and covariance structure of the 'old' distribution. We statistically assessed whether the critical lures fell outside the distribution of the old items by comparing to a threshold of chi^2(2) = 5.99, and found that three of the four critical lures reached the threshold (*ps* < 0.02).

This correlation between spatial and temporal confidence judgments was also present when limiting the analysis only to 'Sure Yes' responses, $r = 0.55$, $p < 0.001$, and the four critical lures again appeared visually pulled away from the distribution. This was again confirmed both with a MANOVA revealing a significant effect of word type, $V = 0.44$, $F(2, 61) = 23.7$, $p < 0.001$ and by comparing the Mahalanobis distance of the critical lures to the 'old' distribution, which again revealed that three of the four critical lures were significantly outside the distribution (*ps* < 0.007; Fig. 5C). Thus, these data suggest that by simultaneously probing memory for multiple contextual features, we can dissociate true and false memories.

5.2. Experiment 5b – In-person sample

To build confidence in the Experiment 5a result (especially given the smaller sample size), we repeated the experiment using an in-lab sample.

5.2.1. Methods

5.2.1.1. Participants. 62 participants were recruited for Experiment 5b to achieve a usable sample size of $N = 50$. Participants were recruited via the subject pool at The Ohio State University and were compensated with course credit.

Given (i) our supplemental analyses suggesting that many of our key results were robust to the strict exclusion criteria (see Supplemental Text); and (ii) that we had greater control over the experimental setting and were able to verify that participants were performing the task in a distraction-free environment, we loosened the inclusion criteria for this experiment. We employed the same criteria as in Experiment 5a, except that our threshold for temporal memory was a Spearman rank correlation >0 (as opposed to ≥ 0.3).

One participant did not false alarm to at least one critical lure; one participant had a memory score > 0 for unrelated foils; one exhibited spatial memory accuracy <0.25; and 9 exhibited a Spearman rank correlation <0.

5.2.1.2. Task design & procedure. The procedure was identical to Experiment 5a, except that participants performed the task in-lab, with an experimenter present to read aloud the instructions at each phase of the task.[1]

[1] This decision had an unintended consequence, which is that there was a relatively large delay between the first word list and the first memory test (as the experimenter read aloud the detailed instructions), and thus participants' memory performance on the first block was significantly worse (and associated with lower spatial and temporal memory confidence) than the other blocks.

At the end of the task, participants were also asked about whether they had pre-experimental familiarity with false memory experiments or the DRM paradigm. Sixteen participants indicated some degree of familiarity (most frequently reporting that they had heard of related studies in their psychology courses). Analyzing the data separately by whether participants had DRM familiarity, we did find that participants who had knowledge of the DRM paradigm were less likely to false alarm to the critical lures, exhibiting a mean memory score of 0.89 for critical lures, compared to an average score of 1.32 in those without prior knowledge, $t(27.33) = 2.15$, $p = 0.041$. However, the key effects (i.e., difference in confidence ratings for critical lures compared to old items) were still observed even in the participants with prior knowledge of the DRM paradigm.

5.2.2. Results

The mean memory score for old words was 1.24 (SD = 0.35), whereas the memory score for unrelated foils was −1.47 (SD = 0.46). The mean memory score for critical lure words was 1.18 (SD = 0.66). Participants were significantly more likely to endorse critical lures as old, compared to unrelated foils, $t(49) = 23.70$, $p < 0.001$, $d = 3.35$. There was no difference in mean memory score for critical lures, compared to old items, $t(49) = 0.70$, $p = 0.48$, $d = 0.10$. Overall, participants exhibited an average accuracy of 44.2% on the spatial memory test (SD = 11.7%) and exhibited an average correlation between their timeline responses and the true temporal positions of *rho* = 0.22 (SD = 0.15).

Considering the reliability in spatial context judgments, we observed greater deviation from chance for old, compared to critical lure items (bootstrap $p < 0.001$); in contrast to Experiment 5a, this additionally held when limiting the analysis only to "Sure Yes" responses (bootstrap $p < 0.001$). However, we still did not replicate the patterns of greater variability in temporal context memory judgments, either for all 'yes' responses (bootstrap $p = 0.66$) or for "Sure Yes" responses only (bootstrap $p = 0.69$), though we note that this is likely because we used a more liberal temporal memory inclusion criteria.

Critically, we replicated the patterns of confidence results from Experiment 5a: Participants were significantly less confident in their spatial (all 'yes' responses: $t(49) = 5.96$, $p < 0.001$, $d = 0.84$; 'Sure Yes' responses: $t(46) = 5.62$, $p < 0.001$, $d = 0.82$; Fig. 5D) and temporal (all 'yes' responses: $t(49) = 3.47$, $p = 0.0011$, $d = 0.49$; 'Sure Yes' responses: $t(46) = 2.78$, $p = 0.0078$, $d = 0.41$; Fig. 5E) memory judgments for critical lures, compared to old items. Further, aggregating confidence ratings for each word across participants, we found reliably lower confidence in both spatial and temporal memory judgments (all bootstrap $ps \leq 0.002$).

Across all words, we observed a strong correlation between spatial and temporal memory confidence, (All 'yes 'responses: $r = 0.88$, $p < 0.001$; 'Sure Yes' only: $r = 0.75$, $p < 0.001$). Considering all 'yes' responses, none of the critical lures were significantly outside of the distribution of the old items (assessed via Mahalanobis distance); however, a one-way MANOVA revealed a significant effect of word type (critical lure vs. old) on the confidence metrics, $V = 0.13$, $F(2, 61) = 4.39$, $p = 0.017$, such that critical lures were low in both spatial and temporal memory confidence. When limiting to 'Sure Yes' responses, one of the four critical lures was significantly outside the 'old' distribution ($p = 0.049$), but again, a one-way MANOVA revealed a significant effect of word type, $V = 0.13$, $F(2, 61) = 4.59$, $p = 0.014$.

We conducted an additional exploratory analysis in which we excluded the first block from the above multi-dimensional analysis; because in the in-lab sample, participants paused between the first word list and first memory test to receive detailed instructions from the experimenter, there was a significantly greater lag between study and test for the first word list, compared to the others. Indeed, confidence on both the spatial and temporal dimensions were significantly lower in block 1, compared to the other 3 blocks ($ps < 0.001$). Across all 'yes' responses, we again found that no critical lure was significantly outside the 'old' distribution, but still observed a significant effect of word type using the MANOVA, $V = 0.15$, $F(2, 45) = 3.82$, $p = 0.029$. For 'Sure Yes' responses, we found that one of three critical lures was significantly outside the 'old' distribution ($p = 0.042$) and again found a significant effect of word type, indicated by the MANOVA $V = 0.19$, $F(2, 45) = 5.43$, $p = 0.0077$ (Fig. 5F).

Together, despite the caveats related to the in-person data collection (i.e., more liberal inclusion threshold and the unanticipated consequences of carefully reading the instructions outloud), these data provide an important validation of the key findings, demonstrating that they are robust in an in-lab sample.

6. General discussion

False memories are not simple errors of recognition, nor are they perfect replicas of true experience. Rather, they appear to occupy an important middle ground — compelling enough to be believed yet subtly impoverished in their contextual grounding. Here, we found that memories for events that never occurred were marked by greater variability for *and* lower confidence in where and when those experiences supposedly took place. These patterns were detectable both across individuals (using variability in context memory across people, as well as mean confidence ratings for individual words) and within individuals (comparing mean confidence ratings within-participant). Critically, these signals emerged even when participants were highly confident that they had encountered the item itself, suggestive of dissociable memory traces for content and source. However, we also identified certain circumstances in which false memories could not as readily be distinguished from true ones. Together, these data highlight novel ways that context representations can (and cannot) be used to draw inferences about true and false memories, opening new avenues for understanding the subjective content of false memories.

6.1. Using context memory to disentangle true and false memories

Across the five experiments, we assessed context memory in two distinct ways. First, we measured *variability* in context memory, assessing the consistency with which people remembered the temporal (and/or spatial) locations of individual words. Second, we measured *confidence* in context memory, asking participants to rate their subjective confidence in their temporal (and/or spatial) memory judgments. These two metrics provide complementary information about whether a memory is true or false; each metric has unique strengths and weaknesses.

For example, the variability approach provides a unique way to measure differences between true and false memories in a maximally open-ended way: Simply by probing memory for spatial and/or temporal position, we can observe a signature of false, compared to true memories. That said, it is also worth noting that this approach is more sensitive to overall data quality, in that the contrast between true and false memories depends on having low variability, or consistent observations for the true memories. Indeed, when employing more liberal inclusion criteria (see Supplemental Text), we were less likely to observe robust differences in variability, likely due to increased noise in the representations of true memories. The confidence approach, in contrast, was more robust to the specific exclusion criteria, though does require that we directly probe participants to reflect on the strength of their subjective experience, which may shape the way participants respond (i.e., causing them to reflect more on whether they did indeed encounter a memory in the first place).

Nevertheless, we found that both of these measures provided a meaningful cue to the veracity of memories at the group level. Specifically, we found that variability in item-level context memory judgments across people (i.e., when/where they remember a word occurring in time/space) was higher for false memories and that aggregated item-level confidence judgments were lower for false memories. Further, in

Experiment 5, we found that by aggregating confidence across both spatial and temporal context judgments, we were able to partially separate true and false memories in a multi-dimensional space. The present work thus provides a template for future investigations to perhaps scale up this approach (i.e., with more data per-participant) and build machine-learning based classifiers to distinguish true and false memories according to these context memory metrics. Additionally, future work could apply this approach to studying false memories outside of the DRM paradigm (Parks et al., 2026): The approach may be particularly fruitful for assessing real-world, collective (false) memories (Maswood & Rajaram, 2019; Prasad & Bainbridge, 2022).

In addition to detecting differences between true and false memories across participants, we also observed reliable within-participant signals using the confidence metric. Specifically, people reported significantly lower confidence ratings in their memory for the contextual features of false, compared to true memories. Importantly, this was not merely the result of decreased confidence in having seen the word: All of our effects held even when limiting the analysis to words that were endorsed as having been seen with the highest level of recognition confidence, suggesting that confidence in context memory can provide additional informative signal, above and beyond mere recognition confidence. Future work could perhaps employ this confidence-based approach to cases where people make context memory errors or misattributions (e.g., Carpenter & Schacter, 2017; Herz et al., 2023): Would people exhibit also lower confidence associated with these more subtle, context misattributions?

It is worth noting that people did tend to have lower recognition confidence for false words as well (i.e., a greater proportion were rated as 'unsure yes', compared to 'sure yes'), replicating Jou et al. (2004). However, it is unclear whether this finding should be taken at face value. On the one hand, it may be true that participants were less confident in having seen the critical lures. Alternatively, participants' confidence in their memory for context could inform their confidence in recognition. That is, although we instructed participants to indicate their recognition confidence irrespective of whether they could remember the context, it is possible that participants used their context memory as a cue to whether they had seen the word. In other words, because we explicitly oriented participants to pay attention to context, they may have adopted a strategy of using context to inform whether (and how confident they are) in having seen a given word to begin with. Such an interpretation would be consistent with findings in the eyewitness testimony (Geiselman et al., 1986) and source monitoring (Johnson et al., 1993; Johnson & Raye, 1981; Lindsay & Johnson, 1989; Unsworth & Brewer, 2010) literatures which suggest that orienting people to contextual aspects of their memory can reduce the formation of false memories. Indeed, insofar as episodic memories are defined as having not just a 'what' component, but also a 'when' and 'where' (Tulving, 1993), orienting participants to whether they remember the 'when' and 'where' may force them to interrogate the extent to which they find their own memories to be genuine.

6.2. Past work on context and false memories

Prior work examining whether false memories contain contextual detail has found conflicting results. For example, Norman and Schacter (1997) found that although people do reliably endorse false memories as having various associated contextual detail (including temporal context information), they are endorsed at lower rates than true memories (see also Mather et al., 1997). These results are perhaps consistent with the current findings: Even in cases in which participants report an overall high confidence in their memory for the spatial or temporal context of false memories, we still observed that those confidence ratings were lower than that of genuine memories.

These findings stand in contrast, however, with other work suggesting that people do have phenomenologically powerful memories for context or source information (e.g., Lampinen et al., 1999; Payne et al., 1996; Read, 1996). Most relevantly, Read (1996) tested participants' memory for the temporal context of each item in a DRM word list and found a reliable bias to remember the critical lure as having occurred in the first part of the list. In contrast, in the current paper, we did not replicate this bias, finding instead that there was a great deal of variability in people's temporal context memory for critical lures. One perhaps critical methodological difference, however, is that we included "filler" unrelated words at the start and end of each list to eliminate primacy and recency effects, which may have limited our ability to detect a systematic bias to remember the critical lure as having occurred toward the start of the list. It is also worth noting that Read simply reported the likelihood of people remembering the critical lure word as having occurred in the first/second/third segment of the list. He did not, however, assess variability in people's position estimates for critical lures vs. old items. It is possible, for instance, that in spite of the systematic bias, there was also significantly greater variability in people's temporal memory estimates for critical lures vs. old items.

6.3. Inducing false memories for context

Although we show that variability in context memory can inform whether a given memory is true or false in most of our experiments, it is also worth noting that there may be some circumstances in which false memories may indeed contain genuine (false) context memory information. Specifically, in Experiment 2, we demonstrated that we could decrease variability in temporal context memory estimates by altering the structure of word presentation at encoding. By presenting the words in a blocked fashion, such that all words associated with one DRM set were presented in the first half of the list and all words associated with a second DRM set were presented in the second half, we observed a reliable bias in people's temporal memory estimates, such that they placed the critical lure in the half of the list that contained that critical lure's associated items. This finding is perhaps in line with prior work on 'content borrowing' which suggests that people can misattribute the features of related memories to the feature of a false memory, leading to increased and more phenomenologically compelling false memories (Abichou et al., 2022; Arndt, 2006; Arndt, 2010; Lampinen et al., 2005; Nieznański et al., 2018). In other words, people can (perhaps strategically) leverage relevant aspects of the encoding experience to make informed guesses about the context of a false memory. However, whether such a result should be interpreted as a genuine false context memory is unclear. On the one hand, the results could be seen as simply the result of an inference based on the semantic context (Arndt, 2006; Hicks & Starns, 2006a, 2006b). On the other hand, others have argued that there can be genuine 'false recollection' when people erroneously retrieve episodic information associated with a genuine event (Parks et al., 2026).

Another relevant example is that of Read (1996), who found a reliable bias to remember critical lures as having occurred early on in a word list. Read interpreted this finding as evidence that the critical lure word can be activated by the associated words early on in the list and reinforced later, leading to a false memory of it having occurred early on in the list; this interpretation is in line with evidence for the 'temporal repetition effect', whereby repeated information is remembered as having occurred earlier in time than non-repeated information (Sherman & Yousif, 2025). Although it is unclear whether and how the temporal repetition effect could produce false memories, it seems plausible that merely activating a given concept (by virtue of seeing associated words) could lead to such a bias effect. Nevertheless, such a question opens up avenues for future research exploring whether it is possible to induce genuine false memories for context that render such memories indistinguishable from true experiences.

6.4. What is a false memory, exactly?

The current findings also address a broader conceptual question: In

what ways do false memories resemble genuine memories, and in what ways do they not? The present data suggest that false memories are neither wholly fabricated experiences nor phenomenologically identical to true memories. On the one hand, participants consistently experienced critical lures as genuinely remembered, often confidently claiming to have seen them. In this sense, false memories appear to recruit many of the same reconstructive and phenomenological processes as true remembering. On the other hand, those memories were largely distinguishable from genuine experiences in the fidelity and confidence of their contextual representations. Thus, the current findings support the view that memories are composed of distinct, dissociable constituents (i.e., content, source) which each contain meaningful signals. A false memory may therefore be best understood as a partial reconstruction — one that is compelling enough to produce a subjective sense of remembering but nevertheless impoverished in other respects. Indeed, our decision to test memory for spatial and temporal context was inspired by the what/when/where framework of episodic memory (Tulving, 1993), which proposes that episodic memories contain not only information about what happened, but also information for the spatiotemporal context of the encoded experience. Our results thus highlight that even when people are confident in having encountered the critical lures, those false memories may not be truly episodic in nature. To this end, our results make contact with frameworks arguing that false memories arise largely from conceptual or semantic, as opposed to episodic, processing (Brainerd et al., 2008; Brainerd & Reyna, 2002; Bransford & Franks, 1971; Gatti et al., 2022; Gatti et al., 2024; Roediger et al., 2001). However, our results also highlight that there are some situations in which false memories can contain contextual information (i.e., Experiment 2), raising questions about whether there may be circumstances in which false memories genuinely recruit episodic processes.

6.5. Limitations of the current study

The insights from the current data are all drawn from the DRM paradigm. Although the DRM paradigm is a common experimental model for assessing false memories, there are several other paradigms for studying false memory. Thus, there is a key open question about the extent to which the current results generalize to other false memory paradigms. This question is particularly pertinent given that different paradigms may be tapping into different kinds of false memories. DRM errors are thought to specifically reflect a type of false memory arising from associative processing (Roediger & Gallo, 2022). These types of errors are distinct from those which arise from 'misinformation effects', in which people falsely remember encountering an event after receiving misleading or suggestive post-event information (e.g., Calvillo & Parong, 2016; Monds et al., 2017; Ost et al., 2013), as well as autobiographical false memories (Patihis et al., 2018; Wilkinson & Hyman Jr, 1998). That said, although it is important to consider that the current results are specific to the DRM paradigm, the fact that DRM errors do not relate to errors on other false memory tasks does not necessarily mean that context-based memory measures would not dissociate true and false memories in those other tasks. Future work can assess whether memory for spatiotemporal context provides a cue to false memory in these other paradigms.

6.6. Conclusion

Remembering is inherently reconstructive. This reconstruction is driven by surrounding context, semantic structure, inference, and prior experience, which collectively 'fill in the gaps'. Yet those reconstructions also leave detectable traces of uncertainty. Some aspects of false memories — such as when and where they supposedly occurred — are not consistently represented with the same fidelity. By leveraging those traces, the current work offers a new framework for determining the boundary between genuine and illusory remembering.

CRediT authorship contribution statement

Brynn E. Sherman: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the Ohio State University IRB. (Approval No. STUDY20251968).

Acknowledgements

I would like to thank Sami Yousif for helpful feedback throughout the project. I am grateful to Chance Mckinnie, Cadence Waller, and Ellie Wintrich for assistance with Experiment 5b data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2026.106782.

Data availability

The data for all experiments are available on our OSF page: https://osf.io/6c3da.

References

Abichou, K., La Corte, V., Bellegarde, A., Nicolas, S., & Piolino, P. (2022). How rich are false memories in a naturalistic context in healthy aging? *Memory, 30*(3), 262–278.
Arndt, J. (2006). Distinctive information and false recognition: The contribution of encoding and retrieval factors. *Journal of Memory and Language, 54*(1), 113–130.
Arndt, J. (2010). The role of memory activation in creating false memories of encoding context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 66.
Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.
Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11*(5), 164–169.
Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic processing in "associative" false memory. *Psychonomic Bulletin & Review, 15*(6), 1035–1053.
Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*(4), 331–350.
Calvillo, D. P., & Parong, J. A. (2016). The misinformation effect is unrelated to the DRM effect with and without a DRM warning. *Memory, 24*(3), 324–333.
Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(3), 335.
Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*(1), 17.
Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38*(7), 833–848.
Gatti, D., Rinaldi, L., Marelli, M., Mazzoni, G., & Vecchi, T. (2022). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General, 151*(2), 363.
Gatti, D., Rinaldi, L., Mazzoni, G., & Vecchi, T. (2024). Semantic and episodic processes differently predict false memories in the DRM task. *Scientific Reports, 14*(1), 256.
Geiselman, R. E., Fisher, R. P., Cohen, G., & Holland, H. (1986). Eyewitness responses to leading and misleading questions under the cognitive interview. *Journal of Police Science and Administration, 14*(1), 31–39.
Herz, N., Bukala, B. R., Kragel, J. E., & Kahana, M. J. (2023). Hippocampal activity predicts contextual misattribution of false memories. *Proceedings of the National Academy of Sciences of the United States of America, 120*(40), Article e2305292120.
Hicks, J. L., & Starns, J. J. (2006a). Remembering source evidence from associatively related items: Explanations from a global matching model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1164.
Hicks, J. L., & Starns, J. J. (2006b). The roles of associative strength and source memorability in the contextualization of false memory. *Journal of Memory and Language, 54*(1), 39–53.

Holmes, J. B., Waters, H. S., & Rajaram, S. (1998). The phenomenology of false memories: Episodic content and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(4), 1026.
Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3.
Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*(1), 67.
Johnson, M. K., & Raye, C. L. (1998). False memories and confabulation. *Trends in Cognitive Sciences, 2*(4), 137–145.
Jou, J. (2008). Recall latencies, confidence, and output positions of true and false memories: Implications for recall and metamemory theories. *Journal of Memory and Language, 58*(4), 1049–1064.
Jou, J. (2011). Conscious and unconscious discriminations between true and false memories. *Consciousness and Cognition, 20*(3), 828–839.
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*(6), 671–686.
Jou, J., Matus, Y. E., Aldridge, J. W., Rogers, D. M., & Zimmerman, R. L. (2004). How similar is false recognition to veridical recognition objectively and subjectively? *Memory & Cognition, 32*(5), 824–840.
Lampinen, J. M., Meier, C. R., Arnal, J. D., & Leding, J. K. (2005). Compelling untruths: Content borrowing and vivid false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 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*(3), 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*(1), 130–135.
Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*(3), 349–358.
Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366.
Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*(12), 720–725.
Maswood, R., & Rajaram, S. (2019). Social transmission of false memory in small groups and large networks. *Topics in Cognitive Science, 11*(4), 687–709.
Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25*(6), 826–837.
Monds, L. A., Paterson, H. M., & Kemp, R. I. (2017). Do emotional stimuli enhance or impede recall relative to neutral stimuli? An investigation of two "false memory" tasks. *Memory, 25*(8), 945–952.
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*(1), 53–71.
Nieznański, M., Obidziński, M., Niedziałkowska, D., & Zyskowska, E. (2018). Context recollection and false memory of critical lures in the Deese/Roediger-McDermott paradigm: The role of encoding-and retrieval-based mechanisms. *Psihologijske Teme, 27*(3), 365–384.
Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25*(6), 838–848.
Ost, J., Blank, H., Davies, J., Jones, G., Lambert, K., & Salmon, K. (2013). False memory≠ false memory: DRM errors are unrelated to the misinformation effect. *PLoS One, 8*(4), Article e57939.
Pardilla-Delgado, E., & Payne, J. D. (2017). The Deese-Roediger-McDermott (DRM) task: A simple cognitive paradigm to investigate false memories in the laboratory. *Journal of Visualized Experiments: JoVE, 119*, Article 54793.
Parks, C. M., Yonelinas, A. P., & Wahlheim, C. N. (2026). Towards a unified theory of false memory for similar episodes. *Psychonomic Bulletin & Review, 33*(2), 88.
Patihis, L., Frenda, S. J., & Loftus, E. F. (2018). False memory tasks do not reliably predict other false memories. *Psychology of Consciousness: Theory, Research and Practice, 5*(2), 140.
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*(2), 261–285.
Pickrell, J. E., McDonald, D. L., Bernstein, D. M., & Loftus, E. F. (2016). Misinformation effect. In *Cognitive illusions* (pp. 406–423). Psychology Press.
Prasad, D., & Bainbridge, W. A. (2022). The visual Mandela effect as evidence for shared and specific false memories across people. *Psychological Science, 33*(12), 1971–1988.
Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review, 3*(1), 105–111.
Roediger, H. L., & Gallo, D. A. (2022). Associative memory illusions. In *Cognitive Illusions* (pp. 402–418). Psychology Press.
Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(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.
Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*(3), 182.
Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467–474.
Sherman, B. E., & Yousif, S. R. (2025). An illusion of time caused by repeated experience. *Psychological Science, 36*(4), 278–295.
Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*(3), 494–500.
Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science, 2*(3), 67–70.
Unsworth, N., & Brewer, G. A. (2010). Individual differences in false recall: A latent variable analysis. *Journal of Memory and Language, 62*(1), 19–34.
Wilkinson, C., & Hyman, I. E., Jr. (1998). Individual differences related to two types of memory errors: Word lists may not generalize to autobiographical memory. *Applied Cognitive Psychology, 12*(7), S29–S46.