Title
Uninvited and unwanted: False memories for words predicted but not seen

Permalink
https://escholarship.org/uc/item/7w22b8gm

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Haeuser, Katja
Kray, Jutta

Publication Date
2022

Peer reviewed

Uninvited and unwanted:
False memories for words predicted but not seen

Katja Haeuser (khaeuser@coli.uni-saarland.de)
Department of Psychology, Saarland University
66123 Saarbruecken, Germany

Jutta Kray (j.kray@mx.uni-saarland.de)
Department of Psychology, Saarland University
66123 Saarbruecken, Germany

Abstract

Previous demonstrations of false memories for predicted but not presented words used slow encoding and immediate retrieval conditions, potentially exacerbating false memory effects. We present two experiments that investigated whether false memories also occur under self-paced encoding and delayed retrieval conditions, and whether false memories are reduced when the initial prediction was disconfirmed by an implausible word, thought to elicit false memory suppression. Results showed that previous demonstrations of false memories were not contingent on the task conditions: False memories also occur when language processing is self-paced, and they affect longer-term memory structures. Crucially, false memories emerged regardless of whether the prediction-disconfirming word was plausible or not. Results are evaluated against a recent psycho-linguistic account that makes diverging predictions regarding the processing consequences of mild and severe violations of plausibility.

Keywords: sentence processing, prediction, reading, false memory

Introduction

In "Foundations of Language" (2002), Ray Jackendoff expressed strong skepticism about whether language processing could be anything but incremental and bottom-up, since top-down prediction of linguistic material would just not be feasible for the language user. Twenty years later, it is clear that language users not only process new material incrementally as it becomes available, but they also generate predictions about the content and form of what they're about to hear or see next. Recent studies have shown that prediction also affects later stages of language processing: When expectations for a specific word are disconfirmed, predictable words do not become suppressed immediately (Rich & Harris, 2021; Rommers & Federmeier, 2018), but cause false memories when probed shortly after sentence encoding (Hubbard et al., 2019). Here, we investigated whether false memory effects also occur for self-paced encoding and delayed retrieval conditions, and whether false memories are more likely to be suppressed when the word that disconfirms the initial prediction is implausible. We begin by reviewing the previous literature on predictability, plausibility, and false memories.

Predictability and plausibility

The predictability of a word is normally quantified by means of the cloze procedure, in which native speakers of a language complete a sentence frame with the first word that comes to mind. Highly predictable words tend to be facilitated during processing, as illustrated in reduced reading rates (Staub, 2015) or decreased N400 ERP components (Van Petten & Luka, 2012).

A concept related to - but distinct from - predictability, is plausibility, i.e., the plausibility with which an event happens in the real world. Plausibility is normally measured in rating studies where participants indicate how plausibly a word completes a sentence. Recently, a number of studies demonstrated that predictability and plausibility have diverging effects on online sentence processing (Brouwer et al., 2021; DeLong & Kutas, 2014; Nieuwland et al., 2019), and potentially, on longer-term memory structures (Kuperberg et al., 2020; see below).

False memories

In classic studies on the false memory paradigm (e.g., Deese, 1959; Roediger & McDermott, 1995), participants are presented with 12-15 study words (e.g., "rest", "bed", "dream" ...) that are close semantic associates of a critical lure (e.g., "sleep"; DRM lists). In subsequent recall and recognition tests, participants show higher rates of false memory for critical lures, compared to new words that were not presented initially. The false memory effect has been shown to be relatively pervasive. For example, false memory judgments are frequently made with high confidence, suggesting that participants have item-specific recollection of the critical lure, as opposed to mere gist-wise familiarity (Yonelinas, 2002). Prominent theories of the false memory illusion (reviewed in Chang & Brainerd, 2021) argue that false memories emerge as a consequence of spreading activation effects during encoding of list words that erroneously activate critical lures, and/or from overt reliance on gist-wise, semantic memory traces during memory retrieval (Roediger et al., 2001).

Recent empirical evidence suggests that false memories also occur for words that are initially predicted during

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

reading, but not actually encountered, at least in conditions when reading is slow and memory for the lure is probed immediately or quickly after following the sentence (Hubbard et al., 2019; Rich & Harris, 2021; Rommers & Federmeier, 2018). Compared to the classic false memory paradigm, where lure words are entrenched through repeated presentation of close semantic associates, the false memory effect found in these newer studies is arguably more powerful, because it is based on a single encounter – or rather, the lack thereof – with the critical lure. Of note, in classic DRM lists, false memories for lures are substantially reduced when studied word lists consist of fewer, rather than more, semantic associates (e.g., four vs sixteen study words; Dodson & Schacter, 2002).

Initial evidence for the existence of prediction-induced false memories was presented in an ERP study by Rommers and Federmeier (2018). In that paper, the authors showed that predictable words (e.g., "hot") that were not actually processed during initial reading (e.g., "Be careful, the top of the stove is very dirty") induced a “pseudo-repetition” effect on the N400 ERP component when they are presented "again" two sentences downstream. In addition to the implicit memory consequences reported in that study, subsequent studies also demonstrated evidence for prediction-related false memories that are more explicit in nature (i.e., emerge in recall or recognition tasks). For example, in a study by Rich and Harris (2021), participants were presented with constraining or neutral sentence contexts that disconfirmed initial predictions, and immediately at sentence offset, were asked whether they had read the predicted word. Reaction times of "no"-responses were slower for words following highly constraining contexts, suggesting that in the high-constraint condition, the lure had been pre-activated during initial reading and lingered in memory.

Another study by Hubbard and colleagues (2019) was divided into several study-test blocks, in which participants first studied a set of constraining sentences, and then were tested on their memory for critical lure words. Lures were predictable words that were not presented during initial reading (e.g., “window” in "Tim threw a rock and broke the camera"). Aggregated results from the recognition test blocks showed that participants were more likely to classify lures as "old", compared to genuinely new nouns. Hence, when predictions are disconfirmed, predicted words are not immediately suppressed, but remain in a somewhat accessible state in memory, at least for a little while.

However, it is unclear whether previous demonstrations of prediction-related false memories may have been contingent on the task conditions inherent to ERP studies that use artificially slow stimulus presentation rates. Such conditions are known to engender unnaturally strong predictability effects during initial reading (Huettig & Guerra, 2019), which may explain the false memory effect later on. In addition, previous studies used relatively short retention intervals, such that lures were mostly probed immediately or shortly after sentence offset.

Hence, one goal in the present study was to investigate whether false memories also emerge when participants control their own pace during reading, and when the lure word is not probed immediately after it has been predicted, but with a delay. This would demonstrate a longer-term false memory effect that extends implicit priming. A second goal of the present study was to test whether prediction-related false memories are more likely to be suppressed when the prediction-disconfirming word is implausible (e.g., “Since Anne is afraid of spiders, she does not like going down into the moon”), potentially forcing re-interpretation of the sentence context. In previous demonstrations of false memory effects, predictable words were disconfirmed by unexpected but plausible words, i.e., words that can still be integrated, leaving the contextual model intact. A recent psycho-linguistic (Kuperberg et al., 2020) links plausibility violations to longer-term memory and learning, arguing that implausible words elicit contextual re-analysis and updating. In relation to false memories, this could suggest that unpredictable-implausible words are likely to stifle or suppress false memories, because they require the language user to abandon their mental representation of prior contextual model (e.g., people might entertain a cartoon scenario where it is somewhat plausible for a person to "go into the moon", see Kuperberg et al., 2020; but see DeLong et al., 2014, for opposing view).

In what follows, we present two studies that aim to replicate the prediction-related false memory effect using self-paced reading and delayed retrieval conditions (Experiment 1), and we investigate whether proportions of false memories are contingent on the plausibility of the word that disconfirmed the initial prediction (Experiment 2). In both studies, participants read sentences silently in a word-by-word self-paced reading task (non-cumulative). Around 15 minutes later, their memory for previously encoded and not encoded nouns (old and new), as well as lure nouns, was probed in a (surprise) recognition test. In order to discriminate between memory judgments resulting from recollection and familiarity, we additionally asked participants to indicate their confidence about their recognition judgments.

Experiment 1

Method

Participants Fifty-three psychology students (34 female, 19 male) with no language and/or neuropsychological disorders between the ages of 21 and 40 participated for course credit. The sample size was motivated by an a-priori sample size estimation using the *WebPower* package in R, assuming a moderate effect size (f=.45), slight sphericity deviation and power of .80. The resulting *n* was 53. Two subjects had to be excluded from further analysis due to a high proportion of abnormally fast RTs (< 50ms) during SPR, resulting in *n* = 51 participants.

Materials We used 44 German sentence frames (taken from Haeuser & Kray, 2021), which strongly constrained expectations towards a predictable noun, e.g., “basement” in “Since Anne is afraid of spiders, she does not like going down into ...”. Half of the sentence frames were completed with an unpredictable (but somewhat plausible) noun that had the same grammatical gender as the predictable noun (e.g., “garden”). Sentence continuations after the noun were added to account for spill-over effects (e.g., “... on her property”). Offline cloze probability ratings showed high cloze probability for predictable nouns (M=.78, SD=.16) and near-zero cloze probability for unpredictable nouns (M=.03, SD=.006). Offline plausibility tests, which asked participants to rate each sentence for its plausibility on a scale from 1 to 7, showed high plausibility for predictable nouns (M=6.57, SD=033) and mild deviations from plausibility for unpredictable nouns (M=4.11, SD=1.33). Predictable and unpredictable nouns were matched in frequency (M=2.74 and M=2.50; $t(85)=1.68$, $p=.1$), based on the Zipf scale from the SUBTLEX DE data base, Brysbaert et al., 2011). They were also matched in word length (M=5.86 and M=6.30; $t(86)=-0.94$, $p=.35$). All sentences were arranged on two experimental lists (with n=44 items each), so that each subject only saw one version of each experimental item. Fifty highly constraining filler sentences from the Potsdam sentence corpus were added to each list in order to make sure that participants continued to make predictions during reading. Yes/No comprehension questions were added for all experimental (i.e. non-filler) sentences to ensure that participants read for comprehension.

Materials for the word recognition (retrieval) task consisted of 88 “old” nouns (i.e., seen during SPR), and 40 “new” nouns (not seen during SPR). In addition, there were 44 “lure” nouns (i.e., nouns previously expected during SPR but not actually seen). All nouns were matched in frequency and length ($F_{\text{Welch}}(2,76.51)=1.23$, $p=.3x$, and $F_{\text{Welch}}(2, 97.47)=2.50$, $p=.09$). They were arranged on two lists, so that each participant saw 106 nouns in total (22 lure, 44 old, 40 new).

Procedure During SPR, participants read sentences on a screen word-by-word (non-cumulative presentation). Each trial started with a fixation cross, presented in the middle of the screen for 300ms, followed by the first word of the sentence. Participants controlled their own speed during reading by pressing the Space bar to reveal the next word while the previous word disappeared. Participants were instructed to read the sentences as quickly as possible and answer the comprehension question as accurately as possible. In the surprise recognition task (which was administered after a 15-minute delay in which participants completed a task of processing speed, not reported on here), participants were instructed to judge whether words were “old” or “new”, additionally indicating their confidence (“sure”, “maybe”). Participants were asked to respond as quickly and accurately as possible by putting the index and middle finger of both hands on the “S”, “D” (sure new, maybe new) and “J”, “K” (maybe old, sure old) bars. At the bottom of every trial, a legend explained the response options (i.e., participants did not need to memorize the meaning of the keys). The experiment was run online using the platform *LabVanced*.

Results

Self-Paced Reading Accuracy on the comprehension questions was high for predictable and unpredictable sentences (M=.93 in both conditions, range: .81-1.00), suggesting that participants were attentive during the experiment and understood the experimental sentences. Raw reading times (RT) were trimmed minimally, by excluding RTs below 100 ms and over 2500 ms. Only RTs from correct trials were included in the analysis. The predictability manipulation affected RTs predominantly in the spill-over region (see Table 1), so we excluded observations from the noun, and aggregated the remaining data over spill 1 and spill 2 words for statistical analysis. Of special interest is whether the unpredictable condition was read more slowly, as this would corroborate observations from the cloze ratings in that a different word was highly expected during reading. Indeed, ANOVA on log-transformed RT showed a significant main effect of condition, $F(1,50)=6.69$, $p=.01$, $\eta2=.001$, showing longer RTs for unpredictable compared to predictable words.

Table 1: Average reading times (± SD) across conditions in Experiment 1 and Experiment 2. PP=predictable-plausible, UP = unpredictable-plausible, UI = unpredictable-implausible.

	Experiment 1			Experiment 2		
	noun	spill1	spill2	noun	spill1	spill2
PP	433 (201)	428 (149)	429 (175)	335 (110)	342 (82)	332 (96)
UP	448 (244)	440 (157)	445 (184)	347 (120)	354 (97)	348 (104)
UI	-	-	-	345 (1127)	384 (122)	370 (115)

Word Recognition On average, the hit rate for “old” nouns (M=.76) was larger than the false alarm rate to “new” nouns (M=.26), suggesting that participants successfully discriminated old and new items. Figure 1 shows the proportion of “old” responses per condition “old” (previously predictable and unpredictable), “new” and “lure” (Hubbard et al., 2019), split out by low and high confidence. For lures and new words, the bars indicate false alarms (i.e., incorrectly endorsing a word as “old”); for old items, the bars represent correct veridical memory for previously seen words. Prior to statistical analysis, we excluded recognition judgments with reaction times larger than twenty seconds, a procedure that removed less than 1% of all data points. One participant was removed from the analysis due to missing data in one cell. A two-by-three

ANOVA revealed a statistically significant interaction between the effects of confidence (levels: low vs high) and condition (levels: lure, new, old-predictable, old-unpredictable), *F*(3,147)=33.07, *p*<.001, *η*2=.17). Follow-up t-tests (Bonferroni-corrected) showed that lures received more "old" judgments than genuinely new nouns overall (low conf.: *t*(49)=2.47, *p*=.1, *d*=.4; high conf.: *t*(49)=7.18, *p*<.001, *d*=.5). In addition, lures received more “old” judgments made with high compared to low confidence (*t*(49)=3.7, *p*<.001). True memory for “old” nouns did not differ between predictable and unpredictable conditions, both *p*’s (low and high conf.) = .10.

Discussion

Experiment 1 showed two main results. First, false memories for predicted (but not actually encountered) words also emerge in paradigms where initial reading is self-paced and where retrieval of lure words is substantially delayed. Hence, previous demonstrations of false memories were not contingent on slow encoding conditions, and false memories affect longer-term memory structures that go well beyond the immediate sentence or discourse context. Second, false memories are more likely to emerge in high- (compared to low-) confidence judgments, which suggests that they instantiate strong feelings of actual recollection (compared to mere familiarity) during retrieval.

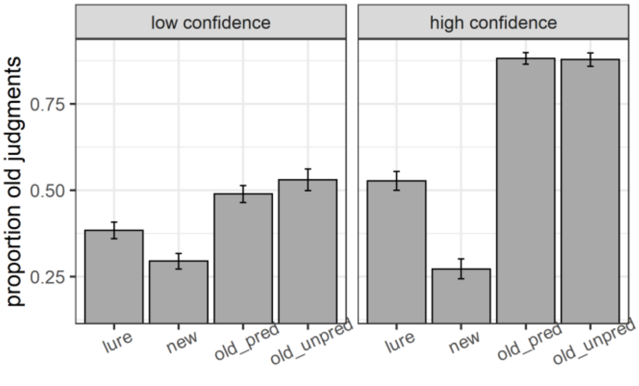

Figure 1: Proportion “old” judgements in the word recognition task from Experiment 1.

In hindsight though, one open question is whether the false memory effect in Experiment 1 may have been driven by the fact that participants’ reading comprehension was probed after each critical sentence. This may have entrenched or re-instantiated critical lures very thoroughly, because not only were lures expected during sentence reading, but they were allowed to linger during the comprehension question phase. Hence, Experiment 2 was set up with two primary goals in mind: First, to shed light on the question whether prediction-related false memories are modulated by the plausibility of the actually presented word, and second, to investigate whether incidental encoding of sentences (with no comprehension question at all) also gives rise to false memories later on.

Experiment 2

Method

Participants Seventy-eight native German speakers (42 female, 35 male, 1 non-binary) with no language and/or neuropsychological disorders between the ages of 18 to 39 (*M*=27 years, *SD*=5) participated for financial compensation (Prolific workers) or course credit (psychology students). This sample size was motivated by an a-priori sample size estimation, assuming an effect size f=.5 (resulting from the condition main effect in Experiment 1, averaged over low and high confidence), slight deviation from sphericity and power of .80. The resulting *n* was 74.

Materials and Procedure The procedure was the same as for Experiment 1. For SPR, we used the predictable and unpredictable sentences from Experiment 1, and added a third condition: unpredictable sentences that were additionally implausible (e.g., “Since Anne is scared of spiders, she does not like going down into the moon ...”). All stimuli were taken from an earlier study conducted in our lab, where unpredictable-implausible nouns were selected on the basis of being impossible in the real world (e.g., it is physically impossible for a person to “go into the moon”). Hence, implausible nouns deviated from the rational assumption that a speaker would communicate literally about possible events in the literal world (Kuperberg et al., 2020). This resulted in three experimental conditions: predictable-plausible (e.g., “basement”, PP), unpredictable-plausible (e.g., “garden”, UP), and unpredictable- implausible (e.g., “moon”, UI). Offline cloze ratings showed that UP and UI nouns were equally unpredictable (M_{UI}=.03, SD_{UI}=.00; *t*(88)=-1.43, p=.2). Pre-study plausibility ratings showed significant plausibility differences between UI items and the other two conditions (PP vs UI: *t*(78)=70.42; UI vs UP: *t*(87)=12.01, both *p*'s < .001). Crucially, UP and UI conditions differed with respect to plausibility: While UP nouns were still somewhat plausible, UI nouns were not. The 132 sentences were evenly distributed on three experimental lists, plus 30 fillers; each subject saw only one version of each experimental sentence. Yes/No comprehension questions were added for filler sentences exclusively.

"Old" nouns for the recognition task were the 88 nouns from Experiment 1, plus 44 UI nouns, resulting in three "old" conditions (PP, UP, UI), and two "lure" conditions: predictable nouns that were previously disconfirmed by unpredictable-plausible nouns (lure-UP), and predictable nouns that were previously disconfirmed by an implausible noun (lure-UI). In addition, there were 50 "new" nouns. Old, new and lure nouns were matched in length and frequency, F_{Welch}(2,102.66)=0.41, *p*=.70, and F_{Welch}(2, 102.59)=1.67, *p*=.21). All nouns were arranged on three experimental lists (on each list: 50 new, 44 old, 44 lure).

Again, SPR and recognition tasks were separated by a 15-minute task that measured processing speed.

Results

Self-Paced Reading Comprehension accuracy for filler items was high across subjects (M=.93, range: .83-1.00). RT outliers below 100ms and above 2500ms were discarded, affecting less than 1% of all data points. The plausibility and predictability manipulations affected reading rates predominantly in the spill-over region, so we excluded observations from the noun, and aggregated the remaining data over spill1 and spill2 words. ANOVA showed a significant main effect of condition on RTs, $F(2,154)=28.18$, $p<.001$, $\eta 2=.01$. Follow-up t-tests (Bonferroni corrected) indicated significant differences for all pairwise contrasts (all p's<.05). Hence, both types of unpredictable-plausible sentences were read more slowly than predictable-plausible ones, and additionally, unpredictable-implausible sentences were read more slowly than unpredictable-plausible ones.

Word Recognition On average, the hit rate for “old” nouns (M=.73) was larger than the false alarm rate to “new” nouns (M=.25). Recognition judgments with RTs>20s were removed. The remaining data were submitted to a six by two ANOVA. The DV was proportion of old judgments. The IVs were confidence (high, low) and condition (old [expected-plausible, unexpected-plausible, unexpected-implausible], new, lure-disconfirmed by unexpected-implausible, lure-disconfirmed by unexpected-plausible). Six participants had missing observations in one or more cells, so they were excluded from the analysis.

Mauchly’s test indicated that the assumption of sphericity had been violated for the interaction between condition and confidence (W=.44, p<.001), therefore, Greenhouse–Geisser corrected tests are reported (epsilon=.77). There was a significant condition by confidence interaction, $F(3.85, 273.35)=29.56$, $p<.001$, $\eta 2 = .11$; see Figure 2). Post-hoc t-tests (Bonferroni-corrected) confirmed the findings from Experiment 1 by showing that, across confidence judgments, participants were more likely to incorrectly endorse lures as previously seen, compared to genuinely new words (high confidence: lure-UI vs new: $t(71)=9.68$, $p<.001$, lure-UP vs new: $t(71)=10.10$, $p<.001$; lure-UP vs new: $t(71)=10.10$, $p<.001$; lure-UP vs new: $t(71)=10.10$, $p<.001$; low confidence: lure-UI vs new: $t(71)=8.12$ $p<.001$; lure-UP vs new: $t(71)=6.97$ $p<.001$). However, unlike in Experiment 1, old judgments to lures did not differ between low- and high-confidence (lure-UP: $t(56)=1.65$, lure-UI: $t(56)=1.51$, both p's >.10). Critical to Experiment 2 is the question whether false alarms to lures differed depending on the plausibility of the prediction-disconfirming word. There were no significant differences between the two types of lures, neither in judgments made with high confidence, $t(71)=.62$, $p=.10$, nor with low confidence, $t(71)=1.23$, $p=.10$. In addition, there were no significant differences with respect to veridical memory for old nouns across confident judgments (all p’s=.10), despite a visible trend for implausible nouns to be remembered more correctly in high-confidence judgments (EMM_{UI}=.90, EMM_{UP}=.85, EMM_{PP}=.84).

Discussion

Three new results emerged from Experiment 2. First, false memories also occur when initial sentence reading is incidental, and when no comprehension question re-instates cognitive representation of the word. Second, in contrast to Experiment 1, where false memory judgments were made with relatively higher confidence, Experiment 2 showed no difference in false memory with regard to confidence (even though there was a numeric trend in this direction). Third, Experiment 2 showed that false memories for predictable words were equally strong, regardless of whether the word that disconfirmed the initial prediction rendered a plausible or implausible reading of the sentence. Thus, despite the fact that implausible sentences may have induced re-interpretation of the contextual model (in fact, the prolonged RTs during sentence encoding suggest that they did), we have no reason to assume that this caused participants to more efficiently suppress critical lures.

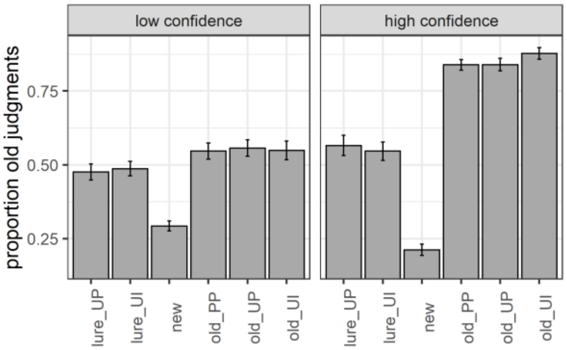

Figure 2: Proportion “old” judgements in the word recognition task from Experiment 2. lure_UP = lure, disconfirmed by unpredictable plausible, lure UI = disconfirmed by unpredictable implausible.

General Discussion

Previous studies showing false memories for predictable words relied on slow stimulus presentation rates during encoding, and on short retention intervals during retrieval, potentially exacerbating false memory effects. We presented two experiments that investigated whether predictable words also elicit false memories when initial reading is self-paced and when retrieval of predictable words is substantially delayed. We also investigated whether false memories are more likely to be suppressed when an initial prediction was disconfirmed by an implausible word that likely triggered re-interpretation of the sentence context. Our results were as follows.

First, results from both experiments demonstrated that false memory effects also emerge when initial sentence encoding is self-paced and when encoding and retrieval are separated by a 15-minute retention interval. Hence, previous demonstrations of false memories were not contingent on slow encoding conditions (Rommers & Federmeier, 2018; Hubbard et al., 2019), and false memories do affect longer-term memory structures, extending beyond the immediate sentence or discourse context. To our knowledge, the two studies reported here are the first behavioral demonstration of prediction-induced false memories using a standard encoding-recognition paradigm (as opposed to multiple encoding-retrieval blocks using only a subset of all items, as in Hubbard et al., 2019).

Second, false memories for predictable words do not seem to "care much" whether the prediction-disconfirming word was plausible or strongly implausible. In Experiment 2, false alarm rates to lure nouns were equally high, irrespective of the plausibility of the prediction-disconfirming word. This suggests that predicted words are not suppressed even when the actually presented word renders an implausible reading of the sentence, potentially triggering active re-structuring or updating of the situation model (Kuperberg et al., 2020).

Third, false memories were more likely to affect high- compared to low-confidence judgments, even though the effect of confidence on lure memory did not reach significance in Experiment 2. Generally, this suggests that recognition judgments for predicted words are not only based on gist-wise familiarity with the item, but on conscious and item-specific recollection. This attests to a certain strength and pervasiveness of the false memory illusion (McDermott & Roediger, 1998).

Finally, the present investigation also showed that false memory effects for critical lures are virtually unchanged even when the lure is not additionally reinstated (and possibly entrenched) by means of a yes/no comprehension question that probes comprehension of the sentence. Hence, false memories also occur when initial reading is incidental. We discuss our findings in greater detail below.

So why was there no plausibility effect for lures in the direction as expected? One possible explanation for the absence of a plausibility effect might lie in the highly constraining nature of our experimental sentences overall. On average, our experimental items were relatively highly constraining towards a specific noun (mean cloze = .77). It is possible that in these conditions, the predictable noun is simply too dominant to allow for effective suppression of the predictable word. For example, we know from studies on lexical ambiguity resolution that, even in conditions when a sentence context strongly biases the subordinate meaning of a homograph (e.g., "bank" of a river), the dominant meaning (e.g., "bank" as in financial institution) becomes activated (e.g., Swinney, 1979; Onifer & Swinney, 1981; Seidenberg et al., 1982). Hence, one possibility is that there should be a more reliable plausibility effect for lures when sentence contexts are only weakly constraining.

A somewhat opposing possibility we need to consider is that the plausibility manipulation did affect recognition judgments for lures after all, but in a way different from what we expected. Notably, there was a small numeric trend in the high-confidence recognition judgment data from Experiment 2, indicating that false memory rates for nouns that were disconfirmed by deeply implausible nouns decreased, whereas correct memory rates for deeply implausible nouns increased. This pattern could indicate a trade-off between true and false memory: As people were more likely to correctly remember deeply implausible nouns as "old", the less likely they were to incorrectly endorse disconfirmed predictable nouns as "old". Ultimately though, it is difficult to draw firm conclusions from these observations, as they mostly rely on (non-significant) patterns in the data. The fact that there was no strong effect for the plausibility manipulation in Experiment 2 is interesting, as it conflicts with a recent psycho-linguistic model of sentence processing (Kuperberg et al., 2020), according to which deeply implausible nouns in highly-constraining sentence contexts that model does not make direct predictions regarding false memories, it does predict that deeply implausible information “entail[s] top-down feedback suppression of the incorrectly predicted semantic features and selection of the correct semantic features” (see p. 24).

The lack of an effect obtained here highlights an open question about the model, specifically, why and how unpredictable, deeply implausible nouns would drive memory and learning over nouns that are mildly implausible, as deeply implausible nouns cannot be integrated meaningfully into a real-world sentence context. Kuperberg and colleagues suggest the possibility of a cartoon scenario that people may entertain in such cases, i.e. language users may temporarily abandon the idea of the rational world to create a fantasy scenario where deeply implausible events could happen. Even though there is evidence for rational adaptation in human cognition (Howes et al., 2009), it seems doubtful whether any kind of adaptation would be that extreme (meaning that language users may still prefer rational context models over irrational ones). Notably, another account of plausibility processing suggests contextual re-analysis (and by extension, a memory advantage) predominantly for nouns that are unexpected but plausible, precisely because they can, to some extent, be integrated into an existing, real-world sentence context (DeLong et al., 2014). Yet others have proposed a memory advantage for deeply implausible words that is based not on language users creating a fantasy scenario, but on a distinctiveness heuristic, i.e. people demand access to distinctive encoding information during retrieval in order to judge an item as old (Dodson & Schacter, 2001). It remains to be seen what type of prediction error (plausible or implausible) is more successful in accounting for longer-term learning and suppressing false memories.

References

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, *12*, 110.

Chang, M., & Brainerd, C. J. (2021). Semantic and phonological false memory: A review of theory and data. *Journal of Memory and Language*, *119*, 104210.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150-162.

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*, 155-161.

Dodson, C. S., & Schacter, D. L. (2002). The cognitive neuropsychology of false memories: Theory and data. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), *Handbook of memory disorders* (pp. 343–362). New York: Wiley.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469-495.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833-848.

Haeuser, K. I., & Kray, J. (2021). Effects of prediction error on episodic memory retrieval: Evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, 1-17.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review*, *116*(4), 717–751.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience*, *13*, 291.

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, *1706*, 196-208.

Jackendoff, R. (2002). *Foundations of language*. New York: Oxford University Press.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12-35.

McDermott, K. B., & Roediger III, H. L. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language*, *39*(3), 508-520.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., ... & Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20180522.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition*, *9*(3), 225-236.

Rich, S., & Harris, J. (2021). Unexpected guests: When disconfirmed predictions linger. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

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

Rommers, J., & Federmeier, K. D. (2018). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage*, *183*, 263-272.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*(4), 489-537.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, *40*(1), 83-108.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*(8), 311-327.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645-659.

Van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211-219.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.