

Are memorability judgements suggestible?

Yanli Wan, Robert A. Nash & Charlotte R. Pennington

To cite this article: Yanli Wan, Robert A. Nash & Charlotte R. Pennington (15 Sep 2025): Are memorability judgements suggestible?, *Memory*, DOI: [10.1080/09658211.2025.2550409](https://doi.org/10.1080/09658211.2025.2550409)

To link to this article: <https://doi.org/10.1080/09658211.2025.2550409>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 15 Sep 2025.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)



This article has been awarded the Centre for Open Science 'Open Data' badge.



This article has been awarded the Centre for Open Science 'Preregistered' badge.

Are memorability judgements suggestible?

Yanli Wan, Robert A. Nash and Charlotte R. Pennington

School of Psychology, Aston University, Birmingham, UK

ABSTRACT

How do we determine whether something that we do not remember actually occurred? People rely partially on judging memorability: when non-remembered events seem memorable, we infer that they did not happen, but when those events seem unmemorable, we might infer instead that they were forgotten. In five online experiments (total $N = 1544$) we examined whether memorability judgements are susceptible to false suggestions. Participants encoded pictures, then completed a test containing old and new pictures. Some test pictures were accompanied by feedback specifying whether they were old or new; however, in a small number of cases, new pictures were falsely identified as “old”. For each picture, participants rated familiarity, subjective memorability, and made judgements of learning. A mega-analysis of Experiments 1–4 showed that participants rated new pictures as less memorable after they received false “old” feedback, compared to no feedback. Moreover, this small feedback effect was stronger for those pictures that people on average found more memorable: a finding replicated in Experiment 5. These findings provide initial empirical evidence that false suggestions, in some circumstances, could subtly shift some people from reasoning “I’d remember this, if it had happened” toward reasoning “I don’t remember this, so maybe it’s forgettable”.

ARTICLE HISTORY

Received 12 March 2025

Accepted 11 August 2025




KEYWORDS

Metamemory; memorability; suggestibility; metacognitive strategy; false feedback

Have you ever taken a hot-air balloon ride, or lost your underwear in the swimming pool? Some readers might readily recall such unusual events, and will easily be able to answer “yes, I have”. But what about the presumed majority of readers who have no immediate memories of these events? Research tells us that people use a *metacognitive strategy* when contemplating the occurrence of non-remembered experiences; that is, they reflect on how memorable the experience would have been if it had happened (Ghetti, 2003; Mazzoni & Kirsch, 2002; Strack & Bless, 1994). If the experience seems highly memorable, then any failure to recall it is usually interpreted as a sign that it never actually happened, and a sign that any suggestion otherwise is clearly false. But if the experience seems highly forgettable, then we might attribute our lack of memory to everyday forgetting and therefore entertain the suggestion that it happened. In short, our judgements of what is memorable can play an important role in shaping our openness to suggestions about past events and experiences. In the present research we asked whether the reverse is also sometimes true: can external influence in the form of false suggestions shape our subjective memorability judgements?

Metamemory and perceived likelihood

Objective memorability, in general, refers to how reliably and consistently an event or item of information is remembered, on average, between people (Bainbridge et al., 2017). Subjective memorability judgements, by extension, concern people’s *predictions about* whether these experiences would be remembered at a later time, and are just one variety of metamemory or metacognitive process through which we monitor and control our memory and cognition (Flavell, 1979; Nelson & Narens, 1990). Subjective memorability judgements may be based on the assumed properties of to-be-remembered material, such as its distinctiveness, or based on people’s own subjective experience of processing the material, such as whether it “feels like” the details have been encoded in memory easily. Whereas these subjective memorability judgements often align with objective memorability, they can also be influenced by heuristics that lead to discrepancies between subjective and objective memorability (Isola et al., 2014). Subjective memorability judgements are also similar to – albeit distinct from – Judgements of Learning (JOLs), which are made during or immediately after learning novel material, and refer to people’s predictions about their future memory performance based on assessments of

CONTACT Robert Nash  r.nash1@aston.ac.uk  School of Psychology, Aston University, Birmingham, B4 7ET, UK
 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09658211.2025.2550409>.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

how well they have learned the material (Nelson & Narens, 1990). Judging subjective memorability may play a role in JOLs, and indeed, people typically give higher JOLs to pictures they consider more rather than less memorable (Saito et al., 2023). However, these two types of judgement are not identical, and some studies even show that JOLs can be a better predictor of objective memorability than are subjective memorability judgements, perhaps because the former rely on more-diagnostic cues (e.g., Navarro-Báez et al., 2025). Importantly, these metacognitive judgements can influence people's behaviour. For instance, when people assign low JOLs to study materials, they often allocate more study time for encoding those items (Mitchum et al., 2016; Nelson & Leonesio, 1988; Thiede & Dunlosky, 1999).

Subjective memorability not only influences people's predictions about future recall, but also plays a crucial role when we make judgements about past events (Ghetti, 2003). Specifically, people can employ metacognitive strategies for deciding the likelihood that an event occurred in the past. In Strack and Bless's (1994) research, participants were presented with various pictures primarily consisting of tools with a few non-tool pictures included. These non-tool pictures were intended to be salient due to their distinctiveness from the other pictures. Later, participants were shown a series of old and new pictures and asked whether they had seen each picture previously. The salience of the test pictures dramatically affected participants' responses: when participants were shown new pictures of non-tools – which were distinctive and therefore memorable – they were very capable of correctly classifying those pictures as “new”. But they struggled much more when shown new pictures of tools, which were less distinctive. In short, these results showed that in circumstances when we would expect to remember, the absence of memory can be a valuable metacognitive cue.

Building upon Strack and Bless's work that involved simple picture stimuli, and the theorising of Koriath and Goldsmith (1996), Mazzoni and Kirsch (2002) considered how people might use a metacognitive strategy when contemplating suggestions about their own past autobiographical experiences. The authors proposed a two-stage model, describing the processes through which people come to believe or disbelieve a suggestion. The first stage involves an effort to retrieve a memory of the suggested event. If a sufficiently clear and vivid memory is retrieved, then this memory directly reinforces the belief that the suggested event occurred. However, if no specific memory is accessed, then the second stage occurs, wherein people draw inferences about the likelihood the event occurred, based on various metacognitive beliefs and information. During this second stage, people first decide whether their lack of memory is “diagnostic” that the event never happened. The authors argue that this diagnosticity decision is based on general metacognitive beliefs about memory as well as specific beliefs about the memorability of the suggested event. As

demonstrated in Strack and Bless's (1994) work, people can easily reject a non-remembered event when their absence of memory seems diagnostic. But when the absence of memory seems not diagnostic, they must instead make inferences based on other available information, such as the plausibility of the suggested event, or the credibility of the suggestion. Mazzoni and Kirsch (2002) propose that only when the apparent likelihood of the event's occurrence exceeds a nominal “belief criterion”, is the suggestion then accepted as true.

Mazzoni and Kirsch's (2002) model has been influential in the literature on autobiographical belief and false memory. Yet the metacognitive process they outlined does not entirely correspond with key findings from the false-memory literature. Specifically, the model proposes that people should easily reject false suggestions about events that they would expect to remember. Numerous studies show, though, that people can be led to remember fictional events that are highly memorable and distinctive. For example, participants in past studies have developed false memories of taking a hot-air balloon ride as a child (Wade et al., 2002) or losing their underwear in the swimming pool (Otgaar et al., 2022). Likewise, participants have been led to falsely remember performing bizarre and highly memorable actions in the very recent past, such as kissing a magnifying glass (Nash et al., 2009; Seamon et al., 2009). Why are these distinctive false suggestions not always easily rejected, as Mazzoni and Kirsch's (2002) model would predict?

One possibility is that even judgements of “diagnosticity” can be susceptible to suggestion. Recall that the Mazzoni and Kirsch (2002) model predicts that people only take external information sources – including false suggestions – into account if (a) they do not remember the suggested event, and (b) they do not judge their non-recollection as being diagnostic that the event never happened. Yet Nash (2009, Nash et al., 2015) proposed that an initial assessment of an event's likelihood might occur *before*, rather than after, judging diagnosticity. The implication is that the same factors which shape people's judgements of event likelihood – such as external evidence and suggestions – could also affect their subjective judgements of how memorable the event would be. In other words, when people encounter a false suggestion about a fictional event that they do not recall, they might reason, “It seems that this happened, so maybe it wouldn't have been so memorable after all”. There is currently a lack of empirical research investigating the hypothesised suggestibility of these metamemory judgements.

The suggestibility of metamemory judgements

Whereas there is limited research directly investigating the suggestibility of subjective memorability, studies have explored the suggestibility of other kinds of metamemory judgement. Memory confidence, for instance, is one type

of metamemory judgement with substantial evidence of its susceptibility to suggestion. Spearing and Wade (2022) had participants watch mock crime videos and complete memory tests, manipulating factors including exposure to misinformation. Participants provided confidence ratings either immediately after each response or at the end of the memory test. The researchers observed that when people were exposed to misinformation, they became overconfident in the accuracy of their memories even when those memories were inaccurate. This overconfidence occurred regardless of the timing of making the confidence judgements, and at almost every confidence level. Other research shows that false suggestions can profoundly influence other forms of metacognitive judgements that underpin belief formation, such as people's beliefs about event plausibility. In one study, researchers showed that childhood events initially perceived as implausible, such as witnessing a demonic possession, could be made to seem plausible through providing carefully crafted narratives and personalised suggestions (Mazzoni et al., 2001; see also Scoboria et al., 2012). This manipulation not only increased participants' ratings of the plausibility of witnessing a possession, but also their ratings of the likelihood it happened during their childhood.

Research has begun exploring how JOLs respond to suggestive influences. For example, Mueller and Dunlosky (2017) manipulated participants' beliefs about how font colour affects processing fluency, telling some participants that one colour (blue or green) was easier to process than the other. Despite no actual differences in recall performance between colours, participants consistently gave higher JOLs to words presented in whichever colour they were led to believe was easier to process. Similarly, Undorf and Bröder (2020, Experiment 3) found that false information had a significant impact on people's immediate JOLs. The authors showed participants a series of word-pairs, and then provided false information that misrepresented the concreteness and emotionality of some of these word-pairs, while providing accurate information for other pairs. The false information influenced immediate JOLs, leading people to be over-optimistic about their likelihood of remembering these pairs in future when concreteness and emotionality were exaggerated, and under-optimistic when concreteness and emotionality were minimised. Results such as these suggest that people can indeed be misled into making faulty predictions about their future memory.

In short, existing research demonstrates the suggestibility of certain metamemory and metacognition judgements that contribute to false beliefs and memories, but we do not yet fully understand whether such misinformation can also shape people's judgements of what is or is not memorable. Particularly, we know little about whether misinformation could lead people to consider it possible they forgot something that (supposedly) already happened.

The present research

Subjective judgements of memorability are critical in equipping people to reject the false suggestions that sometimes cause false memories. Yet even highly memorable events can sometimes be falsely remembered. The present research aimed to test one possible explanation: namely, that false suggestions – in the form of false feedback – can affect people's subjective memorability judgements themselves. Experiment 1 was a pilot study to test for initial evidence in support of this theory. We employed a picture memory paradigm comprising an encoding phase, in which participants saw a series of pictures, and a test phase. In the test phase participants rated the subjective memorability of old and new pictures, sometimes receiving honest feedback about whether they had seen the picture before, and in a small number of cases, receiving false feedback that a truly new picture was actually old.

We examined how this feedback affected participants' JOLs and subjective judgements of memorability, and we predicted that people would judge new pictures as less memorable if they were falsely told they had seen them before. Note that although we included JOLs and subjective memorability judgements as distinct outcome measures, we had no specific *a priori* predictions about how our false feedback might affect these outcomes differentially. Although not crucial to our hypotheses, for exploratory purposes we also examined whether truthful "new" feedback would affect participants' judgements, relative to no feedback. In Experiments 2–4 we then tested the robustness of the effects found in Experiment 1 by replicating this procedure with larger sample sizes, improved methodology, and preregistered design and analysis plans. To synthesise and explore our findings, we then conducted a mega-analysis of the combined data from Experiments 1–4, which indicated that the small effect of false feedback may be larger for pictures that people generally find more memorable. In Experiment 5 we followed up this theoretically important finding in a large and preregistered confirmatory study.

Experiment 1

All five experiments reported in this article received ethical approval from Aston University's Ethics Committee. The experiments were developed in PsychoPy (v23.2; Peirce et al., 2019) and hosted online via the Pavlovia platform in conjunction with Qualtrics. Both the experimental materials and data for Experiment 1 are publicly available on the Open Science Framework: <https://osf.io/w5qk3/>.

Method

Participants

Experiment 1 was a pilot study for which we heuristically chose a target sample size of 60 based upon the sample sizes used in Strack and Bless's (1994) studies. A sensitivity

power analysis using G*Power (v3.1; Faul et al., 2007) indicated that this sample size would allow the detection of a standardised difference between the false-feedback and no-feedback conditions of Cohen's $d_z > 0.37$, assuming 80% power and $\alpha = .05$.

A total of 75 participants were initially recruited through Prolific, of whom 14 were excluded from analysis due to either demonstrating possible awareness of the study hypothesis ($n = 4$), or failing attention/comprehension and data quality checks ($n = 10$). In all the studies described in this paper, the criteria for passing the latter checks included (a) correctly answering specific attention questions, described below, (b) responding on at least 70% of encoding trials, and (c) making identical responses to fewer than 70% of encoding trials. We included criteria (b) and (c) as means of excluding probable inattentive responders, but we acknowledge that in a very small number of cases these criteria may have led to us excluding valid data.

All participants who successfully completed the study and passed the attention checks were paid £2 each. The final sample of $n = 61$ were all English-speaking adults aged 18 or above, and comprised 37 females, 21 males, 1 participant who specified another gender identity, and 2 who chose not to disclose their gender. Participants' mean age was 38.7 ($SD = 11.0$; range = 21–70) with 75% identifying as White, 10% as Black or African American, 7% as Asian, and 8% as belonging to other ethnic categories (note that here and in later parts of the paper, percentages may not sum to exactly 100% due to rounding). A total of 70% of participants were from the UK, 13% from South Africa, 7% from other European countries, 5% from Australia and New Zealand, 3% from Canada and 2% from Mexico.

Materials and procedure

To test our hypotheses, it was important to use stimuli that participants would consider to be reasonably memorable; however, because we planned to use only a relatively mild form of suggestion (i.e., false feedback), it was also important that participants would not find it too difficult to accept that they could forget these same stimuli. We therefore selected 71 photographs of horses from the MemCat database (Goetschalckx & Wagemans, 2019) to use in this study. We opted to use animal pictures based on the reasoning that animals were pre-rated as reasonably memorable in MemCat compared to other superordinate categories such as food and landscapes. We opted to use just one subcategory of animal pictures because we wanted to minimise opportunities for verbal coding in memory (i.e., to avoid the memory benefits of there being one horse, one dog, etc. in the stimulus set), and to enhance the plausibility of participants confusing similar stimuli. We then chose horses somewhat arbitrarily, because the exemplars available in MemCat were visually diverse and emotionally neutral.

Study phase. Participants were informed that they were taking part in an experiment to investigate the memorability of pictures, and began by providing consent and demographic details. Participants were then presented with 50 pictures, each of which appeared sequentially in the centre of the screen for 5 sec. While viewing each picture, participants were prompted to rate the cuteness of the horses ("How cute is this picture?" 1 = *Not at all cute*; 5 = *Extremely cute*). These ratings were not important to our research aims but served to ensure participants' attention toward and encoding of the pictures, as well as providing a potential red herring about the aim of the study. The pictures were displayed in a single pre-determined order for all participants. After the 50 pictures had been presented, participants engaged in a wordsearch puzzle for 5 min, which acted as a filler task.

Test phase. Immediately after completion of the filler task, participants were told they would see a second set of pictures, some of which would be from the set they saw before, and some would be new. They were told that for each picture they would sometimes see feedback about whether or not they had seen the picture before, but sometimes there would be no feedback. Figure 1 visually illustrates the assignment of pictures to different presentation conditions throughout this phase of the study. A total of 30 pictures were presented serially, comprising 9 old pictures that were seen in the study phase and 21 that were new. Of these 30 pictures, ten were shown with no feedback, ten with "old" feedback, and ten with "new" feedback. Specifically, "no feedback" pictures were shown alongside a white box with no additional information. In the "old" feedback condition, pictures were displayed alongside a red box along with the text "You DID see this picture earlier in the study". In the "new" feedback condition, pictures were shown alongside a green box along with the text "You DID NOT see this picture earlier in the study". Within each of these three conditions, four pictures were pre-selected as our "critical" items of interest, all of which were factually new (i.e., they had not been seen before). Of course, this means that in the "old" feedback condition, the feedback for the four critical pictures was false (whereas it was always true for all other critical and non-critical pictures). As a reminder, our central interest was in comparing ratings for the false-feedback and the no-feedback conditions; a true-feedback condition was included for exploratory purposes only.

The 30 test pictures each appeared on the screen until participants responded, without a time limit. To reduce participants' suspicion about the aim of the study, we prevented critical pictures from appearing within the first five test items. The 12 critical pictures were the same for all participants, but their assignment to the three feedback conditions was randomised across participants (four in the no-feedback condition, four in the true-feedback condition, and four in the false-feedback condition). The remaining 18 non-critical pictures, which were largely irrelevant to our analyses, were allocated consistently to the same

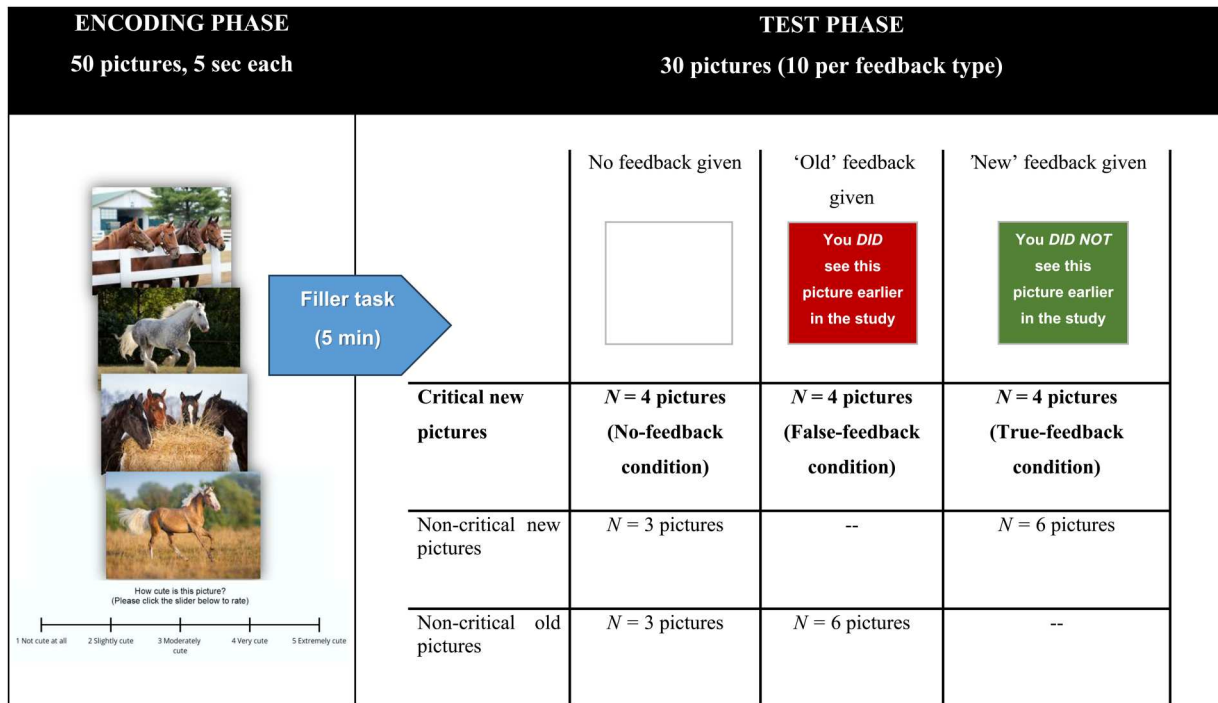


Figure 1. Overview of the experiment design in Experiments 1–4.

conditions across participants: six were truthfully labelled as “old”, six were truthfully labelled as “new”, and the remaining six comprised three old and three new pictures presented without feedback.

When each picture appeared, participants were asked to complete three dependent measures, which we refer to as familiarity, JOL, and memorability (for brevity, we use the term “memorability” here as a shorthand for “subjective memorability”). First, participants were asked “Does this picture seem familiar to you?”, responding on a 6-point Likert scale from 1 (Not familiar at all) to 6 (Extremely familiar). The familiarity rating was not crucial to testing our main hypothesis but served two other functions: it provided auxiliary data to help interpret our main results, and we also anticipated that drawing attention to each picture’s familiarity might boost the predicted effects on the other two dependent variables, because those effects would depend on participants being aware that the critical pictures seem unfamiliar. Participants were then asked the JOL question: “If you saw this picture again in 10 minutes’ time, do you think you would recognise it?”, responding again on a 6-point Likert scale from 1 (I definitely would not recognise this picture) to 6 (I definitely would recognise this picture). Finally, they were asked the memorability question, “How memorable or forgettable do you think this picture is?”, on a scale from 1 (Very forgettable) to 6 (Very memorable).¹ These three questions appeared on the screen underneath each picture and were displayed sequentially, always in the same fixed order. After the three questions had been answered for a given picture, the next picture appeared, and the process repeated.

After all pictures had been rated, an attention/comprehension check was presented, “Which of the following is not a color? (1. Blue, 2. Green, 3. Table, 4. Red)”. This was then followed by two questions to probe participants’ awareness of the study aims: first, participants were asked “What do you think the aim of the study is?” and typed their answer in a text box; second they were presented with a four-alternative multiple-choice version of this question and asked to select which of the following options they believed best represented our research question: (1) Do people believe that simpler images (e.g., a single animal) are easier to memorise than more complex images (e.g., multiple animals)?; (2) Are visually-attractive stimuli more memorable than less-attractive stimuli?; (3) Can people’s perceptions of a picture’s memorability be influenced by false information about whether or not they have seen the picture before?; or (4) Are people better at judging the memorability of a picture if they have actually seen it before?

Results and discussion

Familiarity of old vs. new images

Before addressing our main research question, we examined a feature of our dataset that would evidence participants’ engagement with the task; specifically, we compared the familiarity ratings for the three old versus three new non-critical pictures that were presented in the test phase with no feedback.² A paired sample *t*-test demonstrated that participants rated old non-critical pictures as significantly more familiar ($M = 4.32$, $SD = 1.32$) than new pictures ($M = 2.23$, $SD = 0.86$), $t(60) = 10.61$, p

$< .001$, $d_z = 1.36$. Participants were therefore effective at distinguishing the pictures they saw in the encoding phase from new pictures, indicating their high engagement in the task.

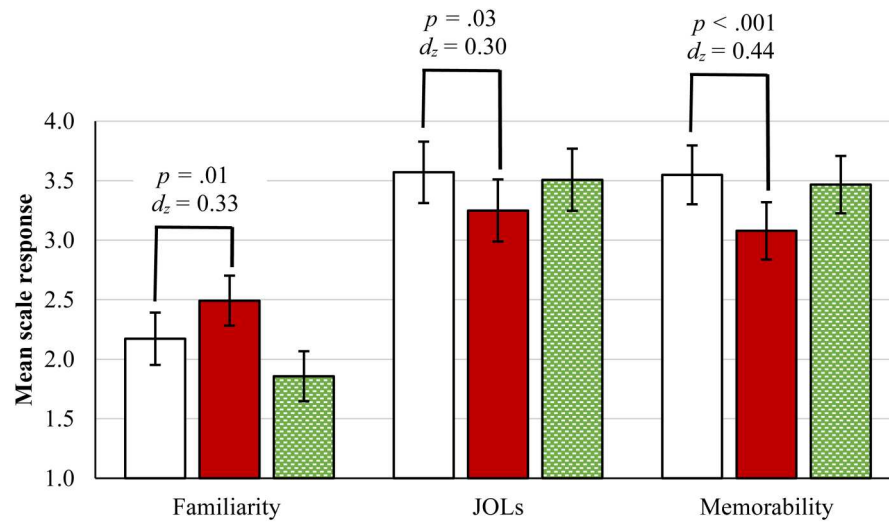
Effects of feedback on metamemory judgements

The means of the primary dependent variables in the three feedback conditions are shown in Figure 2. We first compared the familiarity ratings of pictures across the three conditions. As a reminder, these ratings were not important to test our main hypothesis regarding the impact of false feedback on metamemory judgements, but nevertheless provide a valuable lens through which to interpret the main findings. A series of paired samples t -tests indicated that when participants were falsely told that new critical pictures were “old” (i.e., the false-feedback condition), this led them to rate the pictures as significantly more familiar than no-feedback pictures, $t(60) = 2.59$, $p = .01$, $d_z = 0.33$, and more familiar than pictures shown with true “new” feedback (i.e., the true-feedback condition), $t(60) = 3.95$, $p < .001$, $d_z = 0.51$. Furthermore, participants rated no-feedback pictures as more familiar than true-feedback pictures, $t(60) = 3.05$, $p < .01$, $d_z = 0.39$. These results resemble a classic suggestibility effect, whereby participants’ judgements of familiarity were influenced not only by their memory strength, but also by information they received about their prior exposure to the picture.

Looking to our main hypotheses, our findings for JOLs were different from those for familiarity, and confirmed our predictions. Specifically, receiving false “old” feedback led participants to anticipate being less likely to recognise pictures again in future, as compared with no-feedback pictures, $t(60) = 2.37$, $p = .02$, $d_z = 0.30$, and true-feedback pictures, $t(60) = 2.28$, $p = .03$, $d_z = 0.29$. However, there was no meaningful difference in JOLs between the no-feedback and true-feedback conditions, $t(60) = 0.63$, $p = .53$, $d_z = 0.08$. Our crucial analysis of memorability judgements indicated that false-feedback pictures were judged as significantly less memorable than no-feedback pictures, $t(60) = 3.47$, $p < .001$, $d_z = 0.44$, and true-feedback pictures, $t(60) = 3.13$, $p < .01$, $d_z = 0.40$. There was no meaningful difference in memorability between the no-feedback and true-feedback conditions, $t(60) = 0.85$, $p = .40$, $d_z = 0.11$. Together, these findings from the JOL and memorability measures support our hypothesis that false feedback would mislead people about their likelihood of remembering.

Experiment 2

Experiment 1 provides initial evidence that judgements of subjective memorability can be shaped by misinformation, as proposed by Nash (2009, Nash et al., 2015). However, several methodological limitations of Experiment 1 merit attention before drawing conclusions. First, the experiment involved a relatively small sample of participants,



□ No feedback	$M = 2.17$ $SD = 0.86$	$M = 3.57$ $SD = 1.01$	$M = 3.55$ $SD = 0.96$
■ False feedback	$M = 2.49$ $SD = 1.11$	$M = 3.25$ $SD = 1.13$	$M = 3.08$ $SD = 1.09$
▨ True feedback	$M = 1.86$ $SD = 0.82$	$M = 3.51$ $SD = 1.02$	$M = 3.47$ $SD = 0.94$

Figure 2. Participants’ mean familiarity, JOL, and memorability judgements for critical pictures as a function of feedback condition in Experiment 1 (error bars represent 95% confidence intervals for each individual mean).

with the sample size determined somewhat arbitrarily. Second, the experiment was not preregistered and so a number of researcher degrees of freedom would ideally be removed in a confirmatory study. Third, only 12 critical pictures in the experiment were randomised across the feedback conditions, whereas the remaining non-critical pictures were fixed for all participants. The small number of critical stimuli brings into question whether this effect would replicate with a wider stimulus set. In Experiment 2 we aimed to replicate the findings of Experiment 1 with a preregistered design and analysis plan, a larger sample size, and full randomisation of the entire stimulus set. Specifically, in Experiment 2 the assignment of pictures to the encoding phase and test phase, designation as critical or non-critical, and assignment to different feedback conditions, was determined randomly for each participant. The sequence of the critical and noncritical pictures in the encoding phase and the memory test was also randomised for each participant. Experiment 2 was preregistered on the Open Science Framework prior to data collection: <https://osf.io/rpazn/registrations> and the experimental materials and data are also publicly available: <https://osf.io/rpazn>.

Method

Participants

Based on the results of Experiment 1, we conducted an *a priori* power analysis using G*Power (v3.1; Faul et al., 2007). Our crucial effect rests on the comparison of no-feedback and false-feedback critical pictures, and in Experiment 1 the effect sizes for these comparisons were $d_z = 0.30$ and $d_z = 0.44$ for JOLs and memorability judgements, respectively. Because the small sample size in Experiment 1 could have inflated our estimates of effect size, we decided to power our replication to detect an effect size of $d_z = 0.20$ or larger. A power analysis showed that a sample size of 199 participants would be required for detecting a within-subject difference of this size between the false feedback and no-feedback conditions, assuming $\alpha = .05$ and 80% power.

We initially recruited 239 participants through the Prolific platform who had not participated in Experiment 1, of whom 40 were excluded from the final analysis because of failed attention and data quality checks as described in Experiment 1 ($n = 19$) and/or demonstrating possible awareness of the study's hypotheses based on their response to the open-text awareness check ($n = 21$). The final target of $n = 199$ was therefore achieved. All participants were English-speaking adults aged 18 or above, comprising 132 females, 66 males, and 1 participant who chose not to disclose their gender. Participants' ages ranged from 18 to 81 ($M = 37.3$, $SD = 13.6$) and 72% identified as White, 17% as Black or African American, 5% as Asian, and 6% as belonging to other ethnic categories. A total of 58% were from the UK, 16% from South Africa, 9% from other European countries, 5% from Canada, 4%

from New Zealand, 3% from other African countries, 3% from Israel, 2% from India, and 2% from the USA.

Materials and procedure

The method, materials, and procedure for Experiment 2 was identical to Experiment 1 with the following minor exceptions: all 71 pictures used in Experiment 2 were fully randomised in both the study phase and test phase, as well as to the feedback conditions. Additionally, we removed the requirement that the first five pictures in the testing phase were non-critical.

Results and discussion

Familiarity of old vs. new images

As per Experiment 1, we compared the familiarity ratings for the three old versus three new non-critical pictures in the no-feedback condition. The pictures allocated to each of these groups were randomly selected from the full pool of 71 stimuli, thus removing the possible confound in Experiment 1 where the old and new non-critical pictures were identical for all participants. Here, participants again gave significantly higher familiarity ratings for old pictures ($M = 4.37$, $SD = 1.17$) than for new pictures ($M = 2.31$, $SD = 1.08$), $t(198) = 18.88$, $p < .001$, $d_z = 1.34$, indicating a high level of engagement.

Effects of feedback on metamemory judgements

We next considered participants' familiarity ratings for the critical pictures across the three conditions. Per our preregistration, we carried out three paired-samples *t*-tests. As Figure 3 shows, participants found false-feedback pictures significantly more familiar than no-feedback pictures, $t(198) = 3.54$, $p < .001$, $d_z = 0.25$, and true-feedback pictures, $t(198) = 6.12$, $p < .001$, $d_z = 0.43$. Moreover, participants rated no-feedback pictures more familiar than new-feedback pictures, $t(198) = 3.71$, $p < .001$, $d_z = 0.26$. Consistent with Experiment 1, the pattern of findings resembles a misinformation effect, replicating the well-established power of false suggestions to create illusions of familiarity.

Turning to the JOL data, participants expected themselves less likely to recognise false-feedback pictures in future than to recognise no-feedback pictures, $t(198) = 2.18$, $p = .03$, $d_z = 0.15$, albeit the effect size here was very small, and notably smaller than in Experiment 1. However, there was no significant difference in JOLs between the false-feedback and true-feedback conditions, $t(198) = 1.24$, $p = .22$, $d_z = 0.09$, or between the no-feedback and true-feedback conditions, $t(198) = 1.08$, $p = .28$, $d_z = 0.08$. Finally, looking at memorability judgements, participants rated false-feedback pictures as significantly less memorable compared to no-feedback pictures, $t(198) = 2.64$, $p < .01$, $d_z = 0.19$, and true-feedback pictures, $t(198) = 2.32$, $p = .02$, $d_z = 0.17$, again with small effect sizes. However, subjective memorability did not differ

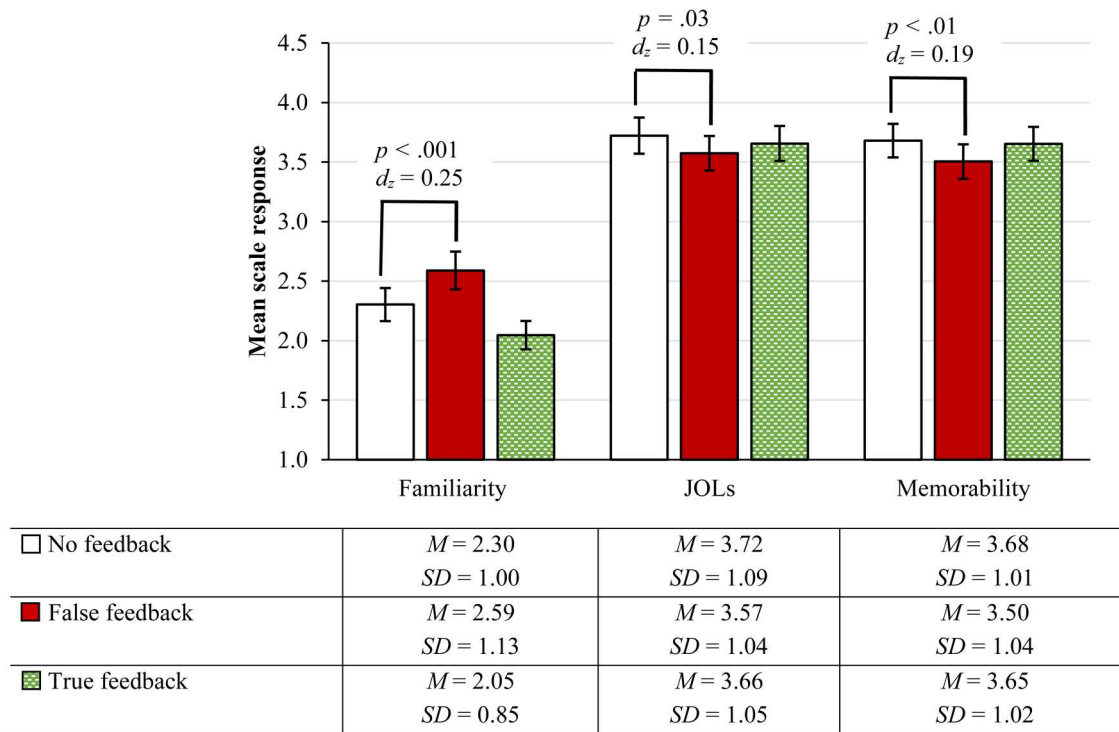


Figure 3. Participants' mean familiarity, JOL, and memorability judgements for critical pictures as a function of feedback condition in Experiment 2 (error bars represent 95% confidence intervals for each individual mean).

significantly between the no-feedback and true-feedback conditions, $t(198) = 0.45$, $p = .66$, $d_z = 0.03$.

In short, we again found that falsely telling people that they had seen a picture before caused them to judge the picture as somewhat less memorable, and to predict themselves somewhat less likely to recognise it in future. Experiment 2 therefore provides the first evidence from a well-powered and preregistered study for the susceptibility of memorability judgements to misinformation. At least two properties of our findings deserve particular attention. One is that effects we observed for familiarity – where false feedback led participants to give higher ratings – fall in the opposite direction to those observed for JOLs and memorability, where false feedback led to lower ratings. Whereas we registered no *a priori* predictions about familiarity effects, these opposing patterns seem to indicate that participants were responding to these questions thoughtfully, and that a complex reasoning process underlies memorability judgements which draws both on metacognitive beliefs and on external evidence (“It appears that I’ve seen this picture before, but even though it seems somewhat familiar I don’t clearly remember it, so maybe it is forgettable”). A second point of note is that our hypothesised effect size was small, and smaller than we had estimated based upon the data from Experiment 1. Indeed, on average the false feedback reduced people’s JOLs and memorability judgements by less than 0.2 points on the respective 6-point rating scales. The small size of this effect does not invalidate its theoretical importance, insofar that our findings lend support to a

theoretical account of how suggestions give rise to false beliefs and memories of unusual events. Nevertheless, drawing conclusions about the applied relevance of this effect demands nuanced interpretation, and we reserve this interpretation until we have outlined further experiments in this paper.

Exploratory analysis

Our randomisation of critical and non-critical stimuli in Experiment 2, whilst enhancing the rigour of the study design, further increased variability between participants in the average memorability of the pictures assigned to each feedback condition. By chance some participants would have seen more-memorable pictures in the “no-feedback” condition than in the “false-feedback” condition and for other participants vice versa. This variability would have added noise to our analysis; therefore, in addition to our pre-registered *t*-test analyses, we replicated these analyses using exploratory linear mixed models to account for this variability.

Specifically, for participants’ JOL ratings, we fitted a linear mixed model using restricted maximum likelihood estimation. The model included feedback condition (no feedback vs. false feedback) as a fixed factor, and random intercepts for participants and for pictures to account for repeated measures and stimulus effects, respectively. For this analysis we ignored the data from true-feedback critical trials because they were not central to our hypotheses. Replicating the results of our main analysis, this mixed-models analysis showed a significant

main effect of feedback condition. Participants gave significantly lower JOLs for false-feedback trials ($M = 3.58$, $SE = 0.08$), than for no-feedback trials ($M = 3.72$, $SE = 0.08$), $M_{diff} = -0.14$, 95% $CI [-0.26, -0.02]$, $t(1371) = 2.36$, $p = .02$. Similarly, for participants' memorability ratings, an equivalent linear mixed model revealed a significant main effect of feedback condition. Memorability ratings were significantly lower in the false-feedback condition ($M = 3.52$, $SE = 0.08$) than the no-feedback condition ($M = 3.67$, $SE = 0.08$), $M_{diff} = -0.16$, 95% $CI [-0.27, -0.04]$, $t(1360) = 2.69$, $p < .01$.³ These results together support our findings that false feedback modestly lowered participants' JOLs and memorability ratings.

Experiments 3 and 4

Even though we detected a significant main effect of false suggestions on subjective memorability judgements in Experiment 2, the effect sizes were notably smaller than in Experiment 1. This difference might reflect normal variability in effect size estimates between individual studies, or might – given that Experiment 2 was more rigorous and better-powered – suggest that the hypothesised effects are weaker than they first seemed. To further establish the reliability and generalisability of our hypothesised effects and to more reliably estimate their size, we conducted two further replications based closely on Experiment 2.

Method

Experiment 3 was identical to Experiment 2 with two exceptions: (1) we once again ensured that the first five pictures shown in the test phase were always non-critical, and (2) we aimed to further enhance data quality by limiting participation to only those people with no prior rejection from other studies on the Prolific platform. For Experiment 3 we followed the same preregistered analysis plan as Experiment 2. Experiment 4 was identical to Experiment 3, except that we attempted to generalise our findings to a new stimulus set from the Memcat database: this time using 71 pictures of hamsters, rather than horses. We preregistered Experiment 4 on the Open Science Framework prior to data collection, noting the change in picture stimuli and also establishing more-explicit criteria for identifying and excluding those participants who might have guessed the study's aims: <https://osf.io/zd9y6/registrations>. All experimental materials and data for Experiment 3 and 4 are publicly available on the OSF: <https://osf.io/g8kwh/> and <https://osf.io/zd9y6/>, respectively.

Participants

In Experiment 3 we recruited 254 participants through the Prolific platform, who had not participated in Experiments 1 or 2. In total, 55 were excluded from analysis due to failing attention checks ($n = 21$), and/or showing possible

awareness of the study hypotheses ($n = 34$). The final sample of $n = 199$ was therefore achieved. All participants were English-speaking adults comprising 120 females, 77 males, 1 participant who identified as another gender, and 1 who did not disclose their gender. Participants' ages ranged from 18 to 77 ($M = 38.5$, $SD = 12.8$) with 74% identifying as White, 14% as Black or African American, 9% as Asian, and 3% reporting another ethnicity. The majority were from the UK (56%), followed by South Africa (13%), Canada (13%), Australia and New Zealand (8%), other European countries (4%), other African countries (4%), other Asian countries (2%), and the USA and Mexico (1%).

Experiment 4 involved 242 English-speaking participants who had not participated in the earlier experiments. In total, 43 were excluded due to failing attention checks ($n = 18$) or exhibiting possible awareness of the study aims ($n = 25$). Again, the final sample size of $n = 199$ was thus achieved, which included 104 females, 91 males, 2 who identified as another gender, and 2 who did not disclose their gender. In this sample, participants' ages ranged from 18 to 77 ($M = 40.5$ years, $SD = 14.1$) with 74% identifying as White, 13% as Black or African American, 6% as Asian, and 8% as other or mixed ethnicity. Half of participants were from the UK (51%), followed by Canada (18%), South Africa (7%), other African countries (7%), United States (6%), other European countries (4%), Australia/New Zealand (4%), Asian countries (2%), and South American countries (1%).

Results and discussion

Familiarity of old vs. new images

In the no-feedback condition of Experiment 3, participants rated old, non-critical pictures as significantly more familiar ($M = 4.37$, $SD = 1.10$) than new, non-critical pictures ($M = 2.03$, $SD = 0.81$), $t(198) = 26.48$, $p < .001$, $d_z = 1.88$.⁴ The same was true in Experiment 4, where old, non-critical pictures were rated as significantly more familiar ($M = 4.41$, $SD = 1.13$) than the new, non-critical pictures ($M = 2.27$, $SD = 1.07$), $t(196) = 19.63$, $p < .001$, $d_z = 1.40$. These large effects again demonstrate that participants were highly engaged with the task, and able to effectively discriminate previously viewed pictures from new ones.

Effects of feedback on metamemory judgements

Figures 4 and 5 show the results of Experiments 3 and 4, respectively. As Figure 4 shows, false-feedback pictures in Experiment 3 were judged as more familiar than both no-feedback pictures, $t(198) = 5.23$, $p < .001$, $d_z = 0.37$, and true-feedback pictures, $t(198) = 6.34$, $p < .001$, $d_z = 0.45$. No-feedback pictures were also judged as more familiar than true-feedback pictures, $t(198) = 2.65$, $p < .01$, $d_z = 0.19$. The same held in Experiment 4: false-feedback pictures were rated as significantly more familiar than no-feedback pictures, $t(198) = 4.05$, $p < .001$, $d_z = 0.29$, and true-feedback pictures, $t(198) = 6.61$, $p < .001$, $d_z = 0.47$,

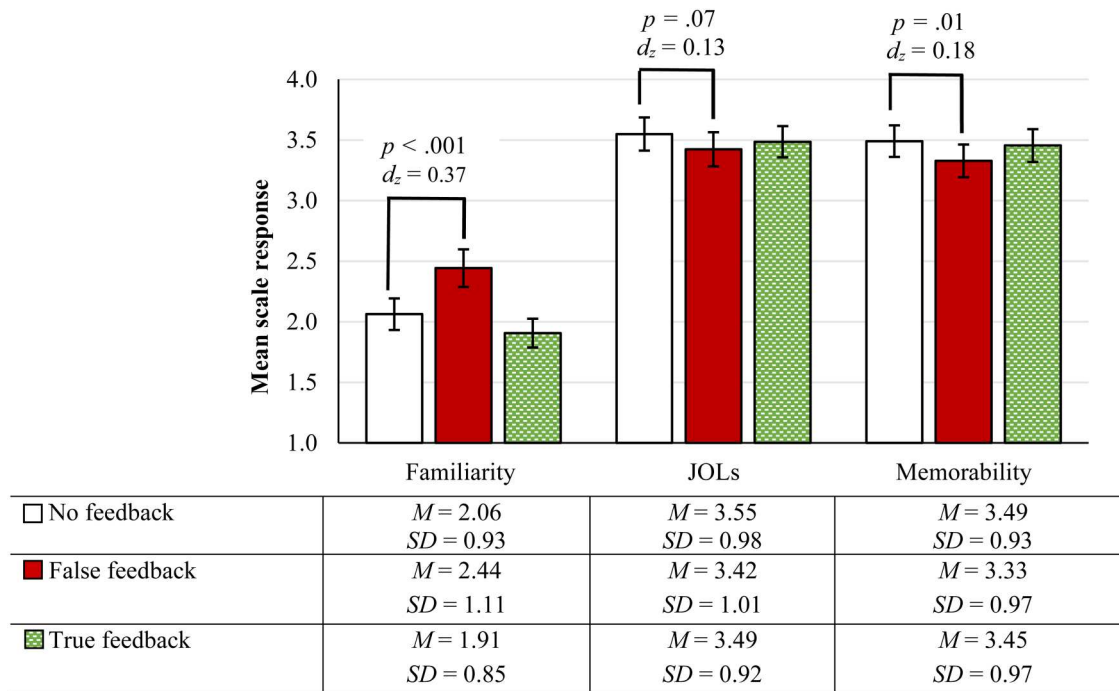


Figure 4. Participants' mean familiarity, JOL, and memorability judgements for critical pictures as a function of feedback condition in Experiment 3 (error bars represent 95% confidence intervals for each individual mean).

and no-feedback pictures were rated as more familiar than true-feedback pictures, $t(198) = 3.05$, $p < .01$, $d_z = 0.22$. These results align with those of Experiments 1 and 2.

However, when looking at JOLs, we found no significant differences across the three feedback conditions in either Experiment 3 or 4. Specifically, in Experiment 3 there

were no significant differences between false-feedback and no-feedback pictures, $t(198) = 1.85$, $p = .07$, $d_z = 0.13$; false-feedback and true-feedback pictures, $t(198) = 0.91$, $p = .37$, $d_z = 0.06$, or no-feedback and true-feedback pictures, $t(198) = 1.02$, $p = .31$, $d_z = 0.07$. In Experiment 4 there were no significant differences in JOLs between

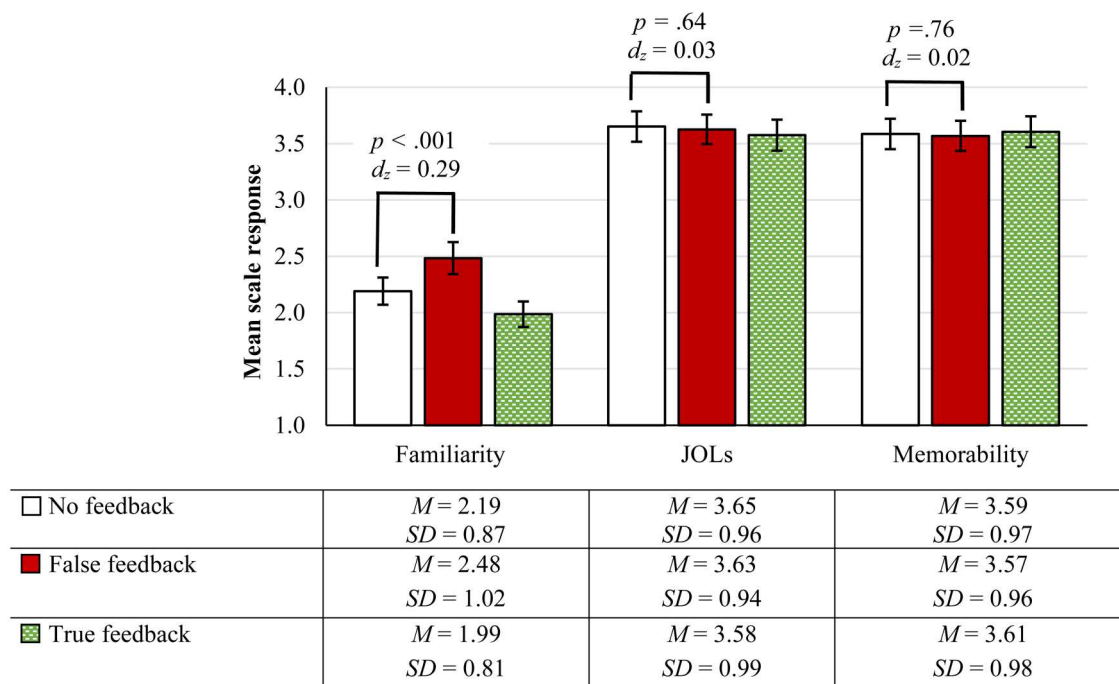


Figure 5. Participants' mean familiarity, JOL, and memorability judgements for critical pictures as a function of feedback condition in Experiment 4 (error bars represent 95% confidence intervals for each individual mean).

false-feedback and no-feedback pictures, $t(198) = 0.47$, $p = .64$, $d_z = 0.03$; false-feedback and true-feedback pictures, $t(198) = 0.96$, $p = .34$, $d_z = 0.07$, or no-feedback and true-feedback pictures, $t(198) = 1.42$, $p = .16$, $d_z = 0.10$. We return to discuss these non-significant findings shortly. For memorability ratings our hypotheses were partially supported. In Experiment 3, as predicted, false-feedback pictures were judged as significantly less memorable than no-feedback pictures, $t(198) = 2.50$, $p = .01$, $d_z = 0.18$, albeit not significantly different from true-feedback pictures, $t(198) = 1.86$, $p = .06$, $d_z = 0.13$, and there was no significant difference between no-feedback and true-feedback pictures, $t(198) = 0.57$, $p = .57$, $d_z = 0.04$. In Experiment 4, there were no significant differences in memorability between feedback conditions: memorability judgements were similar for false-feedback and no-feedback pictures, $t(198) = 0.30$, $p = .76$, $d_z = 0.02$, false-feedback and true-feedback pictures, $t(198) = 0.66$, $p = .51$, $d_z = 0.05$, and for no-feedback and true-feedback pictures, $t(198) = 0.40$, $p = .69$, $d_z = 0.03$.

In sum, the findings of Experiments 3 and 4 provide mixed support for our predictions. Whereas the effects of false feedback were consistently in the predicted direction, the effect sizes we observed were either similar in size to, or even smaller than, our estimates obtained from Experiment 2. The methodology of Experiment 3 was almost identical to Experiment 2; however, Experiment 4 involved a different stimulus set. This difference might therefore lead us to worry that any effect of false feedback is an artefact of the materials used. We return to consider this point further in Experiment 5.

Exploratory analysis

Like in Experiment 2, we conducted linear mixed models analyses for Experiments 3 and 4, which served to test our predictions after taking account of the variability caused by different pictures being assigned to different experimental conditions. In all cases we fitted a linear mixed model using restricted maximum likelihood estimation, with feedback condition (no-feedback vs. false-feedback only, excluding true-feedback critical trials) as a fixed effect, and random intercepts both for participants and for pictures.

Starting with JOLs, in Experiment 3 the analysis demonstrated a significant main effect of condition, which we note is a different conclusion than we drew from our preregistered analysis. Specifically, JOLs were significantly lower for false-feedback pictures ($M = 3.41$, $SE = 0.08$) than for no-feedback pictures ($M = 3.54$, $SE = 0.08$), $M_{\text{diff}} = -0.14$, 95% $CI [-0.25, -0.03]$, $t(1365) = 2.44$, $p = .02$. In Experiment 4, the non-significant effect of false feedback observed in our preregistered analysis was again non-significant in the mixed models analysis: JOLs were similar for false-feedback pictures ($M = 3.59$, $SE = 0.08$) and no-feedback pictures ($M = 3.65$, $SE = 0.08$), $M_{\text{diff}} = -0.06$, 95% $CI [-0.16, 0.05]$, $t(1355) = 1.02$, $p = .31$.

Looking at memorability judgements in Experiment 3, a linear mixed model revealed a significant effect of feedback condition, with memorability rated as significantly lower for false-feedback pictures ($M = 3.31$, $SE = 0.08$) than for no-feedback pictures ($M = 3.49$, $SE = 0.08$), $M_{\text{diff}} = -0.18$, 95% $CI [-0.29, -0.07]$, $t(1359) = 3.14$, $p < .01$. In Experiment 4, this effect of feedback condition was not significant, with little difference in subjective memorability between false-feedback ($M = 3.53$, $SE = 0.09$) and no-feedback pictures ($M = 3.57$, $SE = 0.09$), $M_{\text{diff}} = -0.05$, 95% $CI [-0.15, 0.06]$, $t(1351) = 0.89$, $p = .37$.

These exploratory analyses results bolster those of the preregistered analyses by allowing us to account for the variance in responses caused by participants seeing different critical pictures. The conclusions mirror those of the preregistered analyses, with the exception that they lead us to observe a significant effect of false feedback on JOLs in Experiment 3. Contrary to our hypothesis but consistent with our preregistered analyses, these additional analyses of Experiment 4's data showed non-significant effects of false feedback on both JOLs and memorability ratings.

Mega-analysis of Experiments 1–4

In Experiment 1 we found a significant and modestly sized effect of false suggestions on subjective memorability judgements, whereas in Experiments 2 and 3 this effect was still present but smaller, and in Experiment 4 the effect was almost zero and nonsignificant. To synthesise these effects and afford further insights, we conducted an exploratory (not preregistered) mega-analysis by combining the data across all four experiments. To this end we created a new dataset that contained all the valid data from Experiments 1–4 (total $N = 658$), and then conducted the same analyses on this dataset as for the individual experiments, affording very high power and more-precise estimates of effect size. The data and analysis for the mega-analysis can be found at <https://osf.io/w5qk3/>.

Results and discussion

As expected, the exploratory mega-analysis reproduced the robust effect of false feedback in inflating familiarity judgements: participants rated false-feedback pictures as significantly more familiar than no-feedback pictures, $t(657) = 7.79$, $p < .001$, $d_z = 0.30$, and true-feedback pictures, $t(657) = 11.67$, $p < .001$, $d_z = 0.46$. Moreover, true-feedback pictures were judged as significantly less familiar than no-feedback pictures, $t(657) = 6.09$, $p < .001$, $d_z = 0.24$ (Figure 6).

Looking to our main hypotheses, recall we predicted that false-feedback pictures would be given lower JOLs than no-feedback pictures. This hypothesis was supported in our exploratory mega-analysis, albeit with a very small effect size, $t(657) = 3.37$, $p < .001$, $d_z = 0.13$. Whereas our hypotheses did not specifically concern the true-feedback

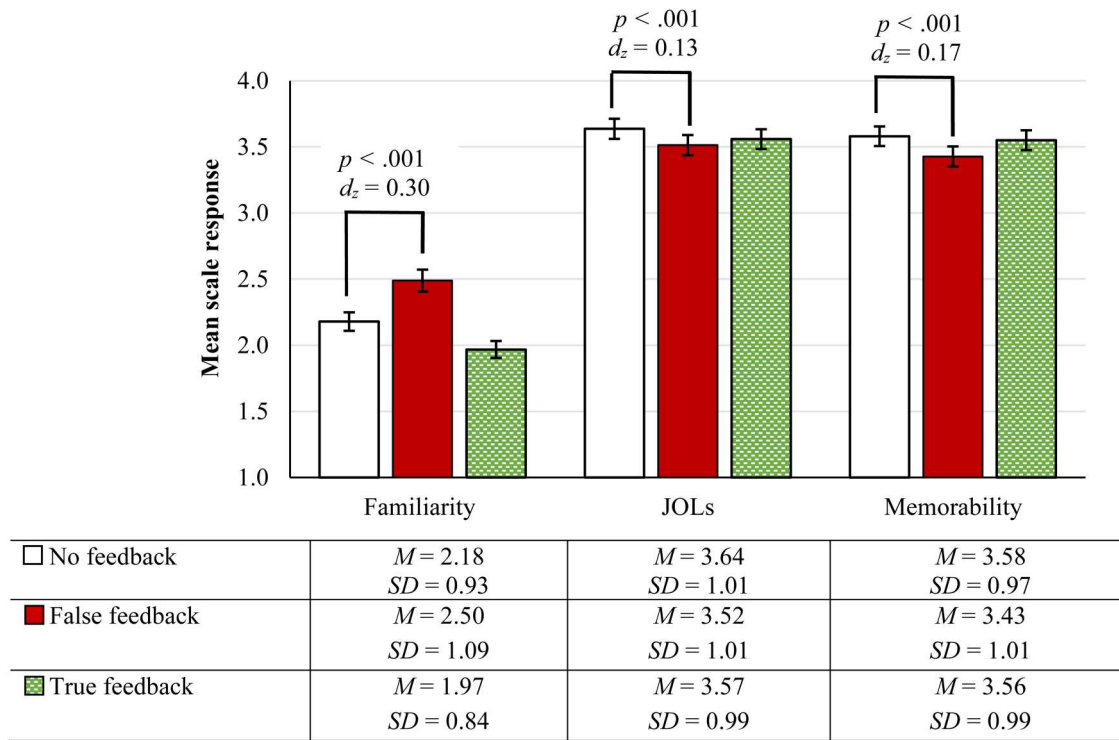


Figure 6. Mean familiarity, JOL and memorability judgements of the critical stimuli as a function of feedback condition in the exploratory mega-analysis of Experiments 1–4 (error bars represent 95% confidence intervals).

condition, we found that JOLs for true-feedback pictures did not significantly differ from JOLs for false-feedback pictures, $t(657) = 1.48$, $p = .14$, $d_z = 0.06$, but were significantly lower – rather than higher, as we might intuitively expect – than JOLs for no-feedback pictures, $t(657) = 2.11$, $p = .04$, $d_z = 0.08$. We return to consider this unexpected exploratory finding in the General Discussion. In terms of memorability judgements, participants rated the false-feedback pictures as significantly less memorable than no-feedback pictures, $t(657) = 4.27$, $p < .001$, $d_z = 0.17$, and true-feedback pictures, $t(657) = 3.73$, $p < .001$, $d_z = 0.15$. There was no difference in memorability ratings between no-feedback and true-feedback pictures, $t(657) = 0.65$, $p = .52$, $d_z = 0.03$.

As in Experiments 2–4, we replicated the exploratory mega-analyses for our key predicted effects using two linear mixed models, to account for the different pictures seen by different participants. A first linear mixed model was fitted using restricted maximum likelihood estimation to examine the effect of false feedback (vs. no feedback) on JOLs, including random intercepts for participants and pictures. The analyses revealed that JOLs were significantly lower for false-feedback pictures ($M = 3.52$, $SE = 0.05$) than for no-feedback pictures ($M = 3.64$, $SE = 0.05$), $M_{\text{diff}} = -0.12$, 95% $CI [-0.18, -0.06]$, $t(4528) = 3.84$, $p < .001$, and that memorability ratings were likewise significantly lower for false-feedback pictures ($M = 3.44$, $SE = 0.05$) than for no-feedback pictures ($M = 3.59$, $SE = 0.05$), $M_{\text{diff}} = -0.15$, 95% $CI [-0.21, -0.09]$, $t(4517) = 4.83$, $p < .001$.

To further quantify the strength of evidence for the false feedback effect, we used jamovi (jamovi Project, 2023) to conduct exploratory Bayesian analyses on this combined dataset, which were not preregistered. For these analyses we used the default Cauchy prior ($r = 0.707$) for two-sided Bayesian t-tests. The results provided strong evidence for the effects of false feedback on memorability judgements ($BF_{10} = 63.32$), showing that the data are over 63 times more likely under the alternative hypothesis than the null hypothesis. For JOLs, the evidence in favour of the alternative hypothesis was substantial but not strong ($BF_{10} = 4.24$). We note, however, that across Experiments 1–4 people's JOLs and subjective memorability ratings were very highly correlated, both in the no-feedback condition ($r = .82$) and the false-feedback condition ($r = .80$).

Moderating effect of agreed subjective memorability

The exploratory mega-analyses described above found, unsurprisingly, that much of the variability in people's JOLs and memorability ratings hinged on the specific pictures they rated. We anticipated our results being of greatest applied and practical relevance if these effects of false feedback generalised to pictures that people considered at least moderately, if not highly memorable (that is to say, if it were possible to reduce the subjective memorability of stimuli that were otherwise considered rather memorable). Therefore, for exploratory purposes we decided to investigate whether the false feedback effect was moderated by

the “agreed” memorability of each unique picture, as assessed as a rating averaged across our participant samples. For every picture stimulus used in Experiments 1–4, we calculated an *agreed memorability* score, i.e., the average memorability rating that had been assigned to that picture by any participant who saw it in the test phase only, in the absence of feedback. For “horse” pictures these agreed memorability scores were therefore based on data from Experiments 1–3, whereas for “hamster” pictures the scores were based on Experiment 4 only. We caution here that the agreed memorability scores are therefore statistically related in part to each individual participant’s memorability ratings. This non-independence of data may introduce statistical artefacts into the results of this exploratory analysis.

We conducted two new exploratory linear mixed models, with JOL and (individual) memorability ratings as the respective dependent variable in each. The models included feedback condition (no-feedback vs. false-feedback), the agreed memorability rating of each unique picture, and the feedback condition \times agreed memorability interaction all as fixed effects, plus random intercepts for participants.

For the JOL model, there was a significant interaction between feedback condition (no-feedback vs. false-feedback) and agreed memorability, $F(1, 4833) = 9.76$, $p < .01$. As Table 1 shows, simple effects analysis found that the effect of false feedback was not significant for low-memorable pictures but was larger for average-memorable and high-memorable pictures. The results of the subjective memorability model mirrored those of the JOL model: there was a significant interaction between feedback condition and agreed memorability, $F(1, 4843) = 16.75$, $p < .001$, with simple effects analyses confirming that the effect of false feedback was not significant for low-memorable pictures, but significant for average-memorable and high-memorable pictures. In short, our exploratory mega-analysis indicated that the effect of false feedback on JOLs and subjective memorability judgements was largest when people were misled about more-memorable, rather than less-memorable pictures.

Experiment 5

To summarise, across our first four experiments combined we found that false feedback had a small yet significant effect on JOLs and subjective judgements of memorability,

but that this effect appeared to be somewhat larger when the feedback pertained to more-memorable pictures. However, our observations about the moderating effect of agreed memorability came from an exploratory analysis, so it is important to replicate this moderating effect in a preregistered, confirmatory study in which agreed memorability is manipulated experimentally. Moreover, when we sought to test the generalisability of the feedback effect to a different stimulus set in Experiment 4, the effect of false feedback was almost zero. This finding might indicate that the false-feedback effect is an artefact of the materials we chose. Alternatively, given the small size of the feedback effect, it is possible that this non-replication in Experiment 4 merely reflected a Type II error: certainly we could foresee no theoretical reason why our effects should occur with pictures of horses and not hamsters. To provide more insight, it remains important to replicate our findings with a more diverse stimulus-set. We designed Experiment 5 to serve both of these aims, making four key adjustments to the basic procedure used in Experiments 1–4.

First, to test the generalisability of our effects we used a new set of pictorial stimuli, depicting visual scenes. Second, we added an experimental manipulation of agreed memorability, ensuring that some of our critical test items were normatively judged as memorable and others judged as unmemorable based on item pre-testing. For clarity, note that we now distinguish two different forms of “memorability” here: we use the term *agreed memorability* to refer to whether, on average, an individual picture was rated as high-memorability or as low-memorability across our pre-test sample of pilot participants. We manipulated agreed memorability as an independent variable in this experiment. In contrast, we use the term *subjective memorability* to refer to each participant’s individual ratings of a picture’s memorability, which was the main dependent variable in this experiment. Third, because we needed a very large participant sample for detecting moderation of an already small effect, we took steps to simplify the study design in ways that would make it more cost-efficient to run at a larger scale. Specifically, we eliminated the true-feedback condition which had been included in Experiments 1–4 for exploratory purposes only, and we removed the JOL measure such that our only crucial dependent variable was subjective memorability. Fourth, to gain deeper insights into participants’ reasoning processes, we added

Table 1. Pictures’ agreed memorability moderated the effect of false feedback upon JOLs and memorability ratings (Experiments 1–4).

Outcome variable	Level of moderator (agreed memorability)	Estimate of mean difference in scores (false feedback – no feedback)	Standard error	df	t	p
JOLs	Low (Mean-1SD)	–0.01	0.04	(4728)	0.23	.82
	Average (Mean = 3.48)	–0.11	0.03	(4603)	3.62	< .001
	High (Mean+1SD)	–0.21	0.04	(4728)	4.76	< .001
Individual memorability ratings	Low (Mean-1SD)	–0.01	0.04	(4734)	0.17	.86
	Average (Mean = 3.48)	–0.14	0.03	(4603)	4.56	< .001
	High (Mean+1SD)	–0.27	0.04	(4734)	6.11	< .001

a written justification phase to collect supplementary qualitative data, described below. This study was preregistered on the Open Science Framework prior to data collection, <https://osf.io/8tz3k/registrations>, and the experimental materials and data can also be found at <https://osf.io/8tz3k/>.

Method

Participants

In our exploratory mega-analysis of Experiments 1–4, the effect of false feedback on participants' memorability judgements was small in size, $d_z = 0.17$. Our exploratory moderation analysis suggested that this effect was larger for pictures that people on average judged more-memorable, but it approached zero for pictures that people on average judged less-memorable.

We planned Experiment 5 not only for detecting the main effect of false feedback on subjective memorability judgements, but also for detecting a hypothesised interaction between feedback and agreed memorability in a 2×2 repeated-measures ANOVA. Given that testing interactions typically requires far larger sample sizes than does testing main effects (Giner-Sorolla et al., 2024), we planned our sample size conservatively, based on being able to detect a false-feedback effect in the high-agreed-memorability condition that was $d_z = 0.17$ larger than the equivalent effect in the low-agreed-memorability condition. We conducted a simulation power analysis using the Superpower shiny app (Lakens & Caldwell, 2021), setting power at 95% and alpha at .05. Based on this simulation we set our target sample size as 886 participants (see our preregistration for a protocol of this power analysis).

We initially recruited 1036 English-speaking adults through the Prolific platform, who had not participated in our previous experiments. After excluding 150 participants for either failing attention and data quality checks as described above ($n = 35$) or showing possible awareness of the study hypotheses ($n = 115$), we met the target sample size of $n = 886$ participants. Of this final sample, 479 were females, 401 were males, 4 identified as another gender, and 2 did not disclose their gender. Participants' ages ranged from 18 to 83 ($M = 37.0$, $SD = 13.1$) with 68% identifying as White, 16% as Black or African American, 11% as Asian, 5% as other ethnicities. Participants were primarily from the UK (54%), Canada (18%), Asian countries (14%), the USA (10%), African countries (8%), other European countries (5%), or Australia and New Zealand (2%). All participants had no prior study rejections in Prolific.

Materials

We first selected an initial set of 150 pictures from the MemCat database (Goetschalckx & Wagemans, 2019), combining landscapes, architecture, and plants, and then recruited 50 participants from Prolific to rate the subjective

memorability of each picture, one at a time, on a scale from 1 (Very forgettable) to 6 (Very memorable). A between-groups split-half reliability analysis demonstrated excellent consistency between independent participants' subjective memorability judgements of these pictures (Spearman-Brown corrected $r = .95$). We then calculated agreed memorability by taking the average rating ascribed to each individual picture across pilot participants, and based on these agreed ratings we selected 94 pictures for use in Experiment 5, of which we selected 40 to be candidate critical pictures (i.e., the 20 pictures with the lowest agreed memorability ratings, and the 20 with the highest ratings). The remaining 54 pictures, whose agreed memorability fell in the mid-range of the scale, were used as non-critical pictures.

Procedure

In the study phase of Experiment 5, we presented 50 pictures to participants, whom we asked to rate the "attractiveness" of each picture. All participants saw the same 50 pictures in a random order. Next, they completed the same 5-minute filler task as used in all previous experiments. In the test phase, participants saw a total of 28 pictures in a random order, 14 of which appeared with the "old" feedback prompt as used in previous experiments, and 14 appeared with no feedback. Of the 28 test pictures, 12 were critical pictures, all of which were always new at test. These 12 critical pictures were subdivided into our four experimental conditions: half were sampled at random from our low-agreed-memorability set, and half from the high-agreed-memorability set, and within each of these two groups, three pictures were selected at random to appear with "old" feedback, and three with no feedback. As a reminder, the "old" feedback was always false for critical pictures. The remaining 16 non-critical pictures comprised 12 "old" pictures from the study phase (8 with true "old" feedback, and 4 with no feedback) and 4 new pictures with no feedback. The 16 non-critical test pictures were always the same for all participants.

At the end of the test phase, participants saw – and rated the familiarity and subjective memorability of – one additional new picture alongside false "old" feedback. Ratings of this picture were not included in our main analyses. Immediately after rating this extra picture, participants were shown the same picture again and told "You rated the memorability of this picture as [X] out of 6. In your own words, please explain the reasoning that led you to choose this rating, why not higher or lower?" This justification task served to look for qualitative evidence of participants consciously using the metacognitive strategy that we have hypothesised.

Results and discussion

Familiarity of old vs. new images

Participants in Experiment 5 rated old, non-critical pictures as significantly more familiar ($M = 5.21$, $SD = 0.90$) than

new, non-critical pictures ($M = 1.80$, $SD = 0.78$), with a very large effect size $t(885) = 79.57$, $p < .001$, $d_z = 2.67$. This finding indicates that participants were highly engaged with the task, although we note that as in Experiment 1 (but not Experiments 2–4), the old and new non-critical items here were not randomised and there is therefore a possibility of item effects.

Effects of feedback on metamemory judgements

As in our previous experiments, we first examined participants' familiarity ratings for the critical pictures in the two conditions. We conducted a 2×2 ANOVA to investigate the effects of feedback (false-feedback vs. no-feedback) and agreed memorability (low vs. high) on familiarity judgements. The results showed a significant main effect of feedback, $F(1, 885) = 58.67$, $p < .001$, $\eta_p^2 = .06$, $d_z = 0.25$. Participants rated false-feedback pictures ($M = 1.86$, $SD = 0.97$) as significantly more familiar than no-feedback pictures ($M = 1.65$, $SD = 0.75$). There was also a significant main effect of agreed memorability, $F(1, 885) = 16.58$, $p < .001$, $\eta_p^2 = 0.02$, $d_z = 0.14$, with low-agreed-memorability pictures ($M = 1.79$, $SD = 0.10$) being rated as significantly more familiar than high-agreed-memorability pictures ($M = 1.71$, $SD = 0.10$). However, there was no significant interaction between feedback and agreed memorability, $F(1, 885) = 0.06$, $p = .81$, $\eta_p^2 < .001$, $d_z = 0.03$.

We tested our primary hypothesis using a two-way repeated measures ANOVA to examine the effects of feedback and agreed memorability on participants' subjective memorability judgements. The results revealed a significant main effect of feedback, with a small effect size, $F(1, 885) = 27.29$, $p < .001$, $\eta_p^2 = .03$, $d_z = 0.18$. As predicted, participants judged false-feedback pictures ($M = 3.11$, $SD = 0.96$) as less memorable than no-feedback pictures ($M = 3.24$, $SD = 0.91$). Confirming the effectiveness of our manipulation, there was also a significant main effect of agreed memorability, $F(1, 885) = 1376.43$, $p < .001$, $\eta_p^2 = .61$, $d_z = 1.25$, with high-agreed-memorability pictures ($M = 3.82$, $SD = 1.05$) being rated as more memorable than low-agreed-memorability pictures ($M = 2.53$, $SD = 0.95$). Crucially, there was a significant interaction between feedback and agreed memorability, $F(1, 885) = 20.91$, $p < .001$, $\eta_p^2 = .02$, $d_z = 0.14$. Post-hoc comparisons using paired t -tests revealed that for high-agreed-memorability pictures, participants made lower subjective memorability ratings in the false-feedback condition ($M = 3.70$, $SD = 1.22$) than in the no-feedback condition ($M = 3.94$, $SD = 1.16$), $t(885) = 6.21$, $p < .001$, $d_z = 0.21$. However, for low-agreed-memorability pictures, participants made similar subjective memorability ratings in the false-feedback ($M = 2.52$, $SD = 1.05$) and no-feedback conditions ($M = 2.55$, $SD = 1.04$), $t(885) = 0.98$, $p = .33$, $d_z = 0.03$ (see Figure 7).

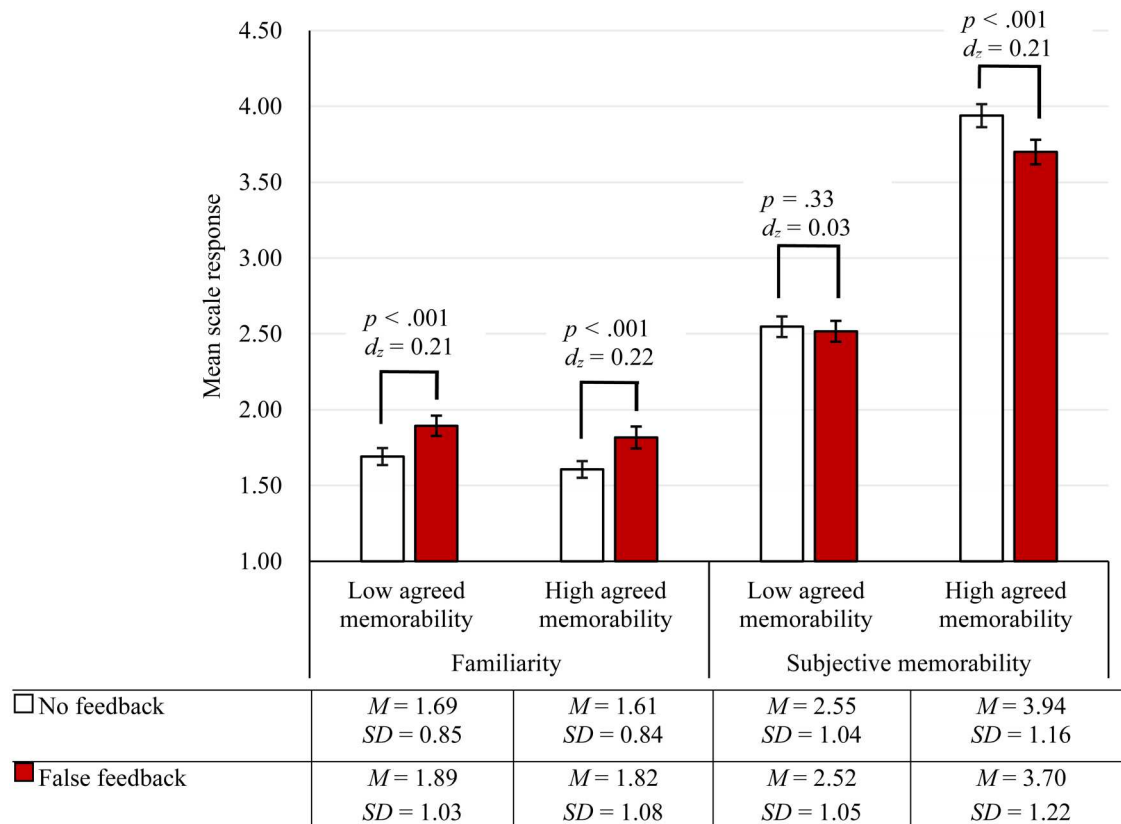


Figure 7. Mean familiarity and subjective memorability judgements of the critical stimuli in Experiment 5, as a function of feedback condition and agreed memorability (error bars represent 95% confidence intervals).

In short, the results of Experiment 5 replicate those of our earlier experiments using a different and more-diverse picture stimulus set. The data also experimentally confirm the exploratory findings from our mega-analysis, in showing that the effect of false feedback on subjective memorability judgements is larger for pictures that were identified in pre-rating data as more memorable. One interpretation of this moderation is that it is an artefact of a floor effect: that is, the subjective memorability of the low-agreed-memorability pictures was already too low to be reduced even further by false-feedback. We find this interpretation unlikely, for two reasons. First, the mean ratings of low-agreed-memorability pictures were substantially above the lower end of the rating scale and, given the absolute size of the effects we observe, there was therefore ample room for feedback to shift people's ratings of these pictures further downwards. Second, we note that in our mega-analysis of Experiments 1–4, the effect of false-feedback was larger for high-agreed-memorability pictures even than for average-agreed-memorability pictures, which points against the moderation effect being attributable to floor effects. Irrespective of the possible role of floor effects, though, our crucial finding here is that the effect of false-feedback persisted even when using more-memorable stimuli.

Exploratory analysis

As in earlier experiments, we also conducted an exploratory linear mixed models analysis, with subjective memorability ratings as the dependent variable. The model included feedback condition (no feedback vs. false feedback), agreed memorability (high vs. low) and their interaction as fixed effects. We also included random intercepts for participants and for pictures. The results revealed that the subjective memorability ratings were significantly lower for false-feedback pictures ($M = 3.11$, $SE = 0.05$) than for no-feedback pictures ($M = 3.24$, $SE = 0.05$), $M_{diff} = -0.13$, 95% CI $[-0.18, -0.08]$, $t(9717) = 5.39$, $p < .001$. We also found a significant effect of agreed memorability, with high-agreed-memorability pictures ($M = 3.82$, $SE = 0.06$) being rated as more memorable than low-agreed-memorability pictures ($M = 2.54$, $SE = 0.06$), $M_{diff} = 1.29$, 95% CI $[1.14, 1.43]$, $t(36.4) = 17.36$, $p < .001$. Finally, we observed a significant interaction between feedback and agreed memorability, $M_{diff} = 0.21$, 95% CI $[0.12, 0.31]$, $t(9716.9) = 4.41$, $p < .001$.

Justifications of reasoning. Recall that we added one “bonus” false-feedback trial at the end of Experiment 5, and asked participants to justify their subjective memorability rating for that picture. The first and second authors independently coded the main theme from each participant's written justification into seven categories (see Table 2 for explanations and illustrative examples of each theme). Initial agreement was 97.0%, and all disagreements were resolved through discussion.

A first, and most crucial, category comprised those participants whose justifications provided unambiguous

evidence of their explicit use of our hypothesised metacognitive strategy. Specifically, these participants directly noted that they were clearly supposed to remember the picture – normally by referencing the feedback they had seen – and yet noted that they did not (clearly) remember it, and that this must mean it is not entirely memorable. In total, eight participants' responses fell into this category. Whereas this is only a tiny fraction of the total participants, the fact that some participants could consciously articulate using this strategy provides convergent evidence to support our interpretations of our experimental findings.

A second category, comprising 11% of participants, referenced their lack of memory as a justification for their subjective memorability rating, but did not explicitly reference a belief that they were supposed to remember the picture. We suspect that some of the people in this category were also consciously using the metacognitive strategy, but simply articulating this without referencing the feedback directly (e.g., “I can't remember seeing this picture before and if I did, it's not very memorable”). In contrast, it is possible that some of the people in this category could have misunderstood the task instructions, by assuming they were rating how well they remembered the pictures, rather than how memorable they were, (for example, the response “I rated it where I did because I did not remember it from the study” might reflect a misunderstanding of the task, or might instead be indirectly referencing how their lack of memory is at odds with the feedback). Other participants coded in this category were clearly referencing their rejection of the false feedback rather than any use of a metacognitive strategy (e.g., “The picture wasn't shown among the images”).

The remainder of participants gave other justifications that were less relevant to our hypotheses, for example by referencing the characteristics or their liking of the picture (61%), contrasting the picture with other pictures (13%), or mentioning personal experiences or self-knowledge of what they are good/bad at remembering (4%). Interestingly, 2% of participants gave ratings of subjective memorability that were based on a belief that the (new) picture was highly familiar or even that they recalled seeing it in the encoding phase. The remaining 8% of participants either gave no response, or gave responses that were ambiguous, irrelevant, or otherwise impossible to code into the other themes.

Exploratory analysis of Experiments 1–5

Recall that in all the five experiments in this paper, we excluded from analysis those participants who may have guessed the study hypothesis based on their responses to the open-text awareness check. As an additional exploratory analysis of interest, we combined all the data from Experiments 1–5 and conducted paired sample *t*-tests to examine the false feedback effect on familiarity and subjective memorability ratings, first for those participants who (in the open-text awareness check) seemed

Table 2. Participants' justifications for their subjective memorability ratings of a picture presented with false "old" feedback in Experiment 5.

Category	Description	Number (%) of participants	Examples
Weak or no memory, despite the feedback	Participants reasoned that they either did not remember the picture, or found it unfamiliar, and yet noted that based on the (false) feedback they should have remembered it	8 (1%)	"I don't remember seeing it but apparently it was shown to me, so I don't think it can be very memorable" "I didn't think I had seen the picture, but the caption indicated I had, leading me to doubt my memory a bit" "It is a beautiful picture, I am surprised that if it was shown I did not remember it. I rated it in the middle, because obviously it should be more memorable, but I do not remember it". "there was a big red box telling me i did see it, i dont remember it therefore it is forgettable"
Weak or no memory, no mention of feedback	Participants reasoned that they either did not remember the picture, or found it unfamiliar, but did not explicitly indicate acceptance of the (false) feedback	97 (11%)	"I like this image so I am surprised that I do not recognise it. It seems quite unique and the colours so stark" "I saw many flower pictures but cannot recall this one. I love pictures of flowers so I am confused as to my recollection" "I haven't seen this picture before" "I don't remember seeing this picture & I paid attention pretty well"
Perceptions and/or liking of the picture's characteristics	Participants' reasoning was based on some visual characteristic of the picture, such as its contents or colouring, or their personal liking of the picture	540 (61%)	"It is a beautiful picture, but not incredibly memorable. It is nice, but also generic". "It is a nice photo but nothing special" "It doesn't stand out in terms of content or colours – it's fairly generic and most likely I wouldn't remember seeing it after a few hours"
Comparison with other pictures	Participants' reasoning was based on a comparison with other pictures, and/or on the picture's perceived similarity with others in the study set.	114 (13%)	"They are just flowers, after seeing many pictures they start to all look the same" "Many flower pictures that are similar" "It is just another picture of a flower among the others, so i dont see it an overly rememebrable compared to them"
Personal experiences or memory abilities/tendencies	Participants either referenced autobiographical experiences that caused the picture to have personal salience, or self-knowledge about their own memory functioning that shaped their expectations about their own likelihood of remembering.	34 (4%)	"because I'm not interested in flowers, so it wouldn't stick out in my mind to remember it. I have a habit of migrating towards things of interest which naturally ill have a higher chance of remembering later on". "Because I remember having a tree like in my neighbourhood growing up" "Flowers are easier to remember for me because I'm a gardener"
Sense of familiarity or false "recollection" of the picture	Participants reasoning was informed by a belief that they had seen the picture before, or that it felt highly familiar	21 (2%)	"I remember this picture and also on the box to the left was written 'you did see this picture'" "Because it felt like I had already seen this picture and the blue color of the flowers is very memorable" "I remembered seeing this picture in the first part of the survey and thought that I would always remember it"
Others	Either no response was provided, or the response was unclear, irrelevant, or otherwise not possible to fit into the other categories.	72 (8%)	"I feel people's feelings are never fully extreme" "FGHFGHFGH" "I am just not too sure about it"

unaware of the study's purpose, and then for those who seemed aware.

Among the group who were unaware of the study's aims, the familiarity ratings were significantly higher for false-feedback pictures compared to no-feedback pictures, $t(1540) = 10.83$, $p < .001$, $d_z = 0.28$, and subjective memorability ratings were significantly lower for false-feedback pictures, $t(1540) = 6.79$, $p < .001$, $d_z = 0.17$. These results reproduce our primary analyses. Among participants who seemed aware of the aims of the study, there was no significant difference in familiarity ratings between the two conditions, and the effect size was notably smaller than

among "unaware" participants, $t(207) = 1.62$, $p = .11$, $d_z = 0.11$. This finding suggests that aware participants may have been able to resist the effects of false feedback on their familiarity judgements, and thus conflicts with a demand-effects account of our findings which would predict larger effects among aware participants. Looking to aware participants' subjective memorability ratings, these ratings were significantly lower in the false-feedback condition than in the no-feedback condition, with an effect size comparable to that among unaware participants, $t(207) = 2.45$, $p < .014$, $d_z = 0.17$. These findings might indicate that even when participants were suspicious of the

feedback, they tended to recognise the deceptive intent to influence their familiarity judgements, but did not recognise their memorability judgements as relevant to the deception. These “aware” participants may therefore still have demonstrated a false-feedback effect on subjective memorability despite their suspicions, perhaps because they did not detect every instance of false-feedback, or perhaps because they retained a grain of doubt that they could be wrong (e.g., “I’m pretty sure they’re lying to me, but if they are not, then this picture might be forgettable after all”).

Finally, we conducted an additional analysis of this combined dataset, in which we also excluded those people who had selected the correct hypothesis from the four-alternative multiple-choice awareness check question. We then ran our main analysis of subjective memorability judgements again on this combined dataset ($N = 853$). The size of the effect of false-feedback on subjective memorability judgements was smaller than in our main analysis, but remained significant, $t(852) = 3.32$, $p < .001$, $d_z = 0.11$. This analysis is highly conservative and likely excludes a large number of unsuspecting participants, but in doing so it ensures a very high likelihood that the remaining participants were unaware of our aims and hypotheses. These exploratory analyses, in combination, therefore point against demand effects being responsible for our findings. More details of this exploratory analysis can be found at <https://osf.io/w5qk3/>.

General discussion

Across five experiments, we aimed to investigate the effects of false suggestions on metamemory judgements. Supporting our main hypothesis, we found that when people were falsely told that they had seen pictures before, which in reality were completely new, people in turn judged those pictures as being somewhat less memorable, albeit the effect of this false suggestion was very small.

Specifically, in Experiments 1, 2, 3, and 5 (but not Experiment 4), false feedback reduced participants’ subjective judgements of the memorability of new pictures, compared to new pictures seen with no feedback. Likewise, a similar pattern of findings emerged for JOLs in Experiments 1, 2, and 5 (but not in Experiment 4, and only in an exploratory mixed-models analysis of Experiment 3). At first glance the null findings in Experiment 4 might reflect methodological differences, as we used a different picture set than in Experiments 1–3. However, Experiment 5, which also used a different picture set, also showed significant effects, thus indicating generalisability of the effect to different stimuli. An alternative interpretation of the mixed pattern of statistical significance across experiments is that this reflects normal variability of a real but small effect, as should be expected in any research that is conducted and reported transparently. Indeed, researchers have pointed out that whenever the proportion of

“significant” results in a series of studies exceeds the statistical power of those studies, this may signify selective reporting rather than robustness of the effects (Cohen, 2013; Vasishth et al., 2018). An exploratory mega-analysis of Experiments 1–4 confirmed a reliable overall cross-experiment effect of false feedback on both subjective memorability judgements and JOLs, yet Bayesian analyses confirmed that the strength of evidence was much greater for the former than for the latter. The weaker findings for JOLs compared to subjective memorability judgements, we speculate, might reflect their distinct underlying cognitive processes, with JOLs often taking into account more diagnostic metacognitive cues (Navarro-Báez et al., 2025). Given the very high correlation between JOLs and subjective memorability ratings, though, this speculation demands further evidence.

Taken together, our findings lend some support to Nash’ (2009, Nash et al., 2015) proposal that false external evidence can shape people’s subjective decisions about the “diagnosticity” of an absent memory. In demonstrating that people can make an initial assessment of an event’s likelihood *before* they judge the diagnosticity of not remembering it, these experiments invite a key theoretical amendment to Mazzoni and Kirsch’s (2002) metacognitive model of false belief. According to their model, when a suggested event is judged to be highly memorable, the absence of a memory for that event normally provides a strong indication that it did not happen, and thus affords easy rejection of the suggestion. In contrast, we show here that false suggestions can undermine this diagnosticity judgement by subtly shaping people’s perceptions of what they should remember well. We emphasise that the effects documented here are very small in size – shifting people’s ratings by on average just a fraction of one point along a six-point rating scale, and that we have focused on picture memory rather than autobiographical memories. Nevertheless, they lend preliminary support to a theoretical account of why people sometimes develop false beliefs and memories about distinctive events, which should be easy to reject by using a memorability heuristic.

Interestingly, our mega-analysis suggested that this false-feedback effect is moderated by pictures’ agreed memorability, with larger effects seen for pictures that people generally considered more memorable. In Experiment 5 we confirmed this pattern of findings in a large, well-powered and preregistered experiment, and we have argued against the likely role of floor effects in accounting for this moderation effect. An alternative interpretation is that the false feedback was more surprising for high-agreed-memorable pictures compared with low-agreed-memorable pictures. Put differently, when people are falsely told they have forgotten a picture that they would not expect to remember anyway, this feedback is unlikely to highlight any cognitive discrepancies that require resolving, and the feedback might therefore receive minimal thought. In contrast, when people are

falsely told they forgot a picture that seems highly memorable, this feedback is likely to create a discrepancy, or expectancy violation, that can only be resolved by either rejecting the feedback, or by recalibrating one's metacognitive judgements (Nelson & Narens, 1990). The exact mechanism behind the apparent relationship between false-feedback and agreed memorability deserves more attention, but perhaps more crucial is our finding that false-feedback affected people's metamemory judgements even when this feedback referred to highly memorable stimuli. To further understand the boundary conditions of this suggestibility effect, future research should extend our work using different and stronger forms of suggestion. To better demonstrate practical relevance, replication studies might also use fewer false-feedback items embedded among larger numbers of non-suggestive items, different types of stimuli or event, and might separate participants' subjective memorability judgements further in time from the false suggestions.

Indeed, the practical relevance of these small effects remains to be demonstrated. For several reasons though, we propose that they could be consequential despite their small size. First, we only used a weak and indirect form of suggestion here, whereas more compelling and powerful suggestions, for example the use of false evidence, typically lead to much stronger suggestibility effects (e.g., Wade et al., 2002). Second, our data indicate that the effect of false feedback was larger when the false suggestions related to more-memorable stimuli. The small effect sizes seen in these studies may therefore underestimate the susceptibility of memorability judgements to suggestion in certain circumstances. Third, even very small effects can have consequences, such as when a witness falsely recalls one single false detail in an otherwise lengthy police report (see Riesthuis et al., 2022). The development of belief and memory errors often begins not with outright acceptance of a falsity, but with a grain of doubt: the entertainment of a possibility that an event *could be* true (Nash et al., 2017). Insofar as we have demonstrated the capacity of false feedback to introduce small doubts about forgettability, these doubts could become the catalyst for larger and more consequential metacognitive errors.

As we have noted, our findings come from a simple picture-memory task with stimuli that hold little personal or emotional relevance, rather than from a task that probes memory for more complex episodic or autobiographical events. A valuable next step in this line of research might therefore be to strive to replicate these effects using more-emotional stimuli, as well as with false-memory paradigms such as Otgaar et al.'s (2022) "blind implantation" technique. If confirmed in these kinds of context, then it would be more convincing to claim that our findings have practical implications, for example in legal, clinical, and therapeutic contexts. Another possible limitation of this study is our sole recruitment of participants from the Prolific platform. Whereas numerous studies have

amassed robust evidence that data quality from Prolific is superior to other online survey platforms (e.g., Amazon Mechanical Turk) and comparable to laboratory samples (Gupta et al., 2021; Palan & Schitter, 2018; Peer et al., 2017), other research suggests that Prolific data can still be "noisier" and exhibit lower responsiveness to certain experimental manipulations, compared to laboratory participants (Gupta et al., 2021; Uittenhove et al., 2023). Attention checks and participant screening can effectively improve data quality, though (Matsuura et al., 2021; Peer et al., 2017), and we implemented several such measures including (1) three attention checks, (2) time constraints for survey completion, and (3) for Experiments 3–5, recruiting only participants with no prior rejections from other studies. These procedures bolster the robustness of our data; nevertheless, replication of these results with populations recruited from other sources would help to establish the generalisability of our findings.

Beyond the main measures and analyses that allowed us to test our main hypotheses, two additional features of our experiments add further insights for interpreting our findings. The first is that participants gave familiarity ratings for each picture. Consistently across each experiment, we found that participants rated new pictures as more familiar after being falsely told they had seen those pictures before, relative to when they were given no feedback. This finding tallies with a literature that is replete with examples of misinformation producing illusions of familiarity, and it indicates that participants frequently accepted the false feedback as valid. However, what is interesting here is that false feedback produced opposing effects on ratings of familiarity – which increased as a result of the suggestion – and on ratings of subjective memorability and JOLs, which decreased as a result of the suggestion. This dissociation between familiarity and memorability/JOL ratings demonstrates that participants were responding thoughtfully, engaging in complex metacognitive reasoning when providing their judgements. They may, we argue, have inferred that if a picture felt familiar despite not being remembered, then it must not be so memorable. Indeed, a very small proportion of the participants in Experiment 5 were able to articulate consciously using this metacognitive strategy.

A second exploratory feature of Experiments 1–4 is that these studies included a "true feedback" condition, wherein participants were truthfully told that certain critical pictures were new. The results from these unregistered analyses are difficult to interpret. On the one hand, our exploratory mega-analysis showed that participants' subjective memorability judgements were unaffected by being told that a new item was new, relative to receiving no feedback. On the other hand, participants' JOLs for true-feedback pictures were equivalent to those for false-feedback pictures, and significantly lower than those for no-feedback pictures. At first glance this latter result seems surprising, especially given that true "new" feedback did (as one might expect) lead people to provide

lower ratings of picture familiarity. We should be wary of drawing theoretical conclusions from non-registered analyses, but one possible explanation is that the (true) “new” feedback provided a different metacognitive cue for some participants than did the (false) “old” feedback. A picture that has been seen twice should, logically, be more recognisable if seen again in future than a picture seen only once. The “new” feedback – which confirmed a picture as being shown for the first rather than the second time – could therefore have given some participants a probabilistic basis for assigning lower JOLs. These findings should alert us that a shift in metacognitive judgements could indicate multiple different or competing metacognitive strategies; more research designed to gain direct insight into those strategies, as we have attempted via our justification task in Experiment 5, would therefore be valuable for this reason. We also presume that people’s likelihood of using these metacognitive strategies might be associated with other stable individual differences. To advance this line of research, future studies should investigate individual differences that shape people’s use of these strategies, and their susceptibility to believing an unremembered experience could be forgettable. One relevant individual difference could be individuals’ chronic levels of trust or distrust in their own memory abilities (Nash et al., 2023; Squire et al., 1979).

Recall that none of the participants included in our analyses showed explicit awareness of the study’s hypothesis when we asked them to guess, and that those people who did seem aware (and who were therefore excluded from our main analyses) actually showed signs of trying to resist being influenced, rather than efforts to confirm the study’s hypotheses. Moreover, across experiments the effect of false feedback on subjective memorability remained significant (although still very small) even among those participants who failed to identify our hypothesis from a four-alternative forced-choice test. These design features help us to discount the likelihood of our findings being attributable to demand effects. A similar counterexplanation of our findings is that participants were reluctant to appear forgetful, and so gave lower stimulus memorability ratings after false feedback as a way of justifying to the researchers – rather than to themselves – why they had failed to remember those stimuli. However, given that participants in this study were never asked to report whether or not they actually remembered seeing each picture, it seems unlikely that this form of social desirability would explain the results we observed.

In sum, the current research provides initial evidence in support of the theory that subjective memorability judgements are suggestible, albeit the average effect size seen across these studies is small. Our findings help to advance existing theoretical models of belief formation, by highlighting the potential role of external influences in shaping judgements of “diagnosticity”. Our findings – if replicated in other contexts with greater ecological validity – could shed light on the cognitive mechanisms

that underlie the development of false beliefs and memories for memorable events.

Transparency & openness statement

All experimental materials and data are publicly available on the Open Science Framework. Experiment 1: <https://osf.io/w5qk3/>; Experiment 2: <https://osf.io/rpazn/>; Experiment 3: <https://osf.io/g8kwh/>; Experiment 4: <https://osf.io/zd9y6/> and Experiment 5: <https://osf.io/8tz3k/>. The associated preregistrations for specific experiments in this article can also be found on the OSF, Experiment 2: <https://osf.io/rpazn/registrations>; Experiment 4: <https://osf.io/zd9y6/registrations>; and Experiment 5: <https://osf.io/8tz3k/registrations>.

Notes

1. The JOL question is overtly framed as a confidence-based measure. However, it is plausible that whereas some participants would also understand our memorability question as a confidence-based measure (i.e., “how confident are you that you would remember this?”), others would understand it as a normative measure (i.e., “would most people tend to remember this?”). It would be useful for future research to differentiate these interpretations, and we thank an anonymous reviewer for this suggestion.
2. Note that in Experiment 1, the old and new non-critical pictures were not counterbalanced across participants, potentially introducing item effects. We addressed this limitation in subsequent experiments.
3. We repeated these linear mixed models by including random slopes for participants, and the pattern and significance of all results was the same as reported here. The same was true when we added random slopes for participants to all of the linear mixed models that we report for Experiments 3–5, and our mega-analysis. The results of these additional analyses can be found in the supplementary materials.
4. After collecting data for Experiment 3 we discovered a programming error that meant the number of non-critical pictures assigned to each of the feedback conditions was not consistent between participants as intended. The assignment of critical pictures to feedback conditions was unaffected by this error, which was corrected in Experiment 4.

Open Scholarship



This article has earned the [Center for Open Science](#) badges for Open Data, Open Materials and Preregistered+. The data and materials are openly accessible at <https://osf.io/w5qk3/>;

Disclosure statement

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

References

- Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from

- memory. *NeuroImage*, 149, 141–152. <https://doi.org/10.1016/j.neuroimage.2017.01.063>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press. <https://doi.org/10.4324/9780203771587>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/bf03193146>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Ghetti, S. (2003). Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language*, 48(4), 722–739. [https://doi.org/10.1016/S0749-596X\(03\)00005-6](https://doi.org/10.1016/S0749-596X(03)00005-6)
- Giner-Sorolla, R., Montoya, A. K., Reifman, A., Carpenter, T., Lewis, N. A., Aberson, C. L., Bostyn, D. H., Conrique, B. G., Ng, B. W., Schoemann, A. M., & Soderberg, C. (2024). Power to detect what? Considerations for planning and evaluating sample size. *Personality and Social Psychology Review*, 28(3), 276–301. <https://doi.org/10.1177/10888683241228328>
- Goetschalckx, L., & Wagemans, J. (2019). Memcat: A new category-based image set quantified on memorability. *PeerJ*, 7, e8169. <https://doi.org/10.7717/peerj.8169>
- Gupta, N., Rigotti, L., & Wilson, A. (2021). The experimenters' dilemma: Inferential preferences over populations. *arXiv:2107.05064*. <https://arxiv.org/abs/2107.05064>
- Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1469–1482. <https://doi.org/10.1109/TPAMI.2013.200>
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103(3), 490–517. <https://doi.org/10.1037/0033-295X.103.3.490>
- Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920951503>
- Matsuura, T., Hasegawa, A. A., Akiyama, M., & Mori, T. (2021). Careless participants are essential for our phishing study: Understanding the impact of screening methods. In *Proceedings – EuroUSEC 2021: 2021 European symposium on usable security* (pp. 36–47). (ACM international conference proceeding series). Association for Computing Machinery. <https://doi.org/10.1145/3481357.3481515>
- Mazzoni, G., & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 121–145). Cambridge University Press. <https://doi.org/10.1017/CBO9780511489976.007>
- Mazzoni, G. A. L., Loftus, E. F., & Kirsch, I. (2001). Changing beliefs about implausible autobiographical events: A little plausibility goes a long way. *Journal of Experimental Psychology: Applied*, 7(1), 51–59. <https://doi.org/10.1037/1076-898X.7.1.51>
- Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General*, 145(2), 200–219. <https://doi.org/10.1037/a0039923>
- Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language*, 93, 245–258. <https://doi.org/10.1016/j.jml.2016.10.008>
- Nash, R. A. (2009). *The metacognitive roles of external evidence in memory construction* [Doctoral dissertation, University of Warwick]. <http://webcat.warwick.ac.uk/record=b2317891~59>
- Nash, R. A., Saraiva, R. B., & Hope, L. (2023). Who doesn't believe their memories? Development and validation of a new Memory Distrust Scale. *Journal of Applied Research in Memory and Cognition*, 12(3), 401–411. <https://doi.org/10.1037/mac0000061>
- Nash, R. A., Wade, K. A., Garry, M., Loftus, E. F., & Ost, J. (2017). Misrepresentations and flawed logic about the prevalence of false memories. *Applied Cognitive Psychology*, 31(1), 31–33. <https://doi.org/10.1002/acp.3265>
- Nash, R. A., Wade, K. A., & Lindsay, D. S. (2009). Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory & Cognition*, 37(4), 414–424. <https://doi.org/10.3758/MC.37.4.414>
- Nash, R. A., Wheeler, R. L., & Hope, L. (2015). On the persuadability of memory: Is changing people's memories no more than changing their minds? *British Journal of Psychology*, 106(2), 308–326. <https://doi.org/10.1111/bjop.12074>
- Navarro-Báez, S., Undorf, M., & Bröder, A. (2025). Predicting the memorability of scene pictures: Improved accuracy through one's own experience. *Quarterly Journal of Experimental Psychology*, 78(3), 546–565. <https://doi.org/10.1177/17470218241239829>
- Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the “labor-in-vain effect”. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14(4), 676–686. <https://doi.org/10.1037/0278-7393.14.4.676>
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Otgaar, H., Moldoveanu, G., Melis, V., & Howe, M. L. (2022). A new method to implant false autobiographical memories: Blind implantation. *Journal of Applied Research in Memory and Cognition*, 11(4), 580–586. <https://doi.org/10.1037/mac0000028>
- Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, 36(1), 203–215. <https://doi.org/10.1002/acp.3911>
- Saito, J. M., Kolisnyk, M., & Fukuda, K. (2023). Judgments of learning reveal conscious access to stimulus memorability. *Psychonomic Bulletin & Review*, 30(1), 317–330. <https://doi.org/10.3758/s13423-022-02166-1>
- Scoboria, A., Mazzoni, G., Jarry, J., & Shapero, D. (2012). Implausibility inhibits but does not eliminate false autobiographical beliefs. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(4), 259–267. <https://doi.org/10.1037/a0030017>
- Seamon, J. G., Blumenson, C. N., Karp, S. R., Perl, J. J., Rindlaub, L. A., & Speisman, B. B. (2009). Did we see someone shake hands with a fire hydrant?: Collaborative recall affects false recollections from a campus walk. *The American Journal of Psychology*, 122(2), 235–247. <https://doi.org/10.2307/27784394>
- Spearing, E. R., & Wade, K. A. (2022). Providing eyewitness confidence judgments during versus after eyewitness interviews does not affect the confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, 11(1), 54–65. <https://doi.org/10.1037/h0101868>
- Squire, L. R., Wetzel, C. D., & Slater, P. C. (1979). Memory complaint after electroconvulsive therapy: Assessment with a new self-rating instrument. *Biological Psychiatry*, 14(5), 791–801.
- Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language*, 33(2), 203–217. <https://doi.org/10.1006/jmla.1994.1010>

- The jamovi Project. (2023). *jamovi* (Version 2.4). [Computer software]. <https://www.jamovi.org>
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 1024–1037. <https://doi.org/10.1037/0278-7393.25.4.1024>
- Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, 6(1), 13. <https://doi.org/10.5334/joc.2596+4785>
- Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgments is strategic. *Quarterly Journal of Experimental Psychology*, 73(4), 629–642. <https://doi.org/10.1177/1747021819882308>
- Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103, 151–175. <https://doi.org/10.1016/j.jml.2018.07.004>
- Wade, K. A., Garry, M., Read, J. D., & Lindsay, D. S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review*, 9(3), 597–603. <https://doi.org/10.3758/bf03196318>