Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–20

DOI: 10.1177/17470218241239829
qjep.sagepub.com
Sage

Predicting the memorability of scene pictures: Improved accuracy through one's own experience

Sofia Navarro-Báez[1,2], Monika Undorf[2] and Arndt Bröder[1]

Abstract

There are conflicting findings regarding the accuracy of metamemory for scene pictures. Judgements of stimulus memorability in general (*memorability judgements* [MJs]) have been reported to be unpredictive of actual image memorability. However, other studies have found that *judgements of learning* (JOLs)—predictions of one's own later memory performance for recently studied items—are moderately predictive of people's own actual recognition memory for pictures. The current study directly compared the relative accuracy and cue basis of JOLs and MJs for scene pictures. In Experiments 1 and 2, participants completed an MJ task and a JOL task in counterbalanced order. In the MJ task, they judged the general memorability of each picture. In the JOL task, they studied pictures and made JOLs during a learning phase, followed by a recognition memory test. Results showed that MJs were predictive of general scene memorability and relied on the same cues as JOLs, but MJ accuracy considerably improved after the JOL task. Experiment 3 demonstrated that prior learning experiences drove this increase in MJ accuracy. This work demonstrates that people can predict not only their own future memory performance for scene pictures with moderate accuracy but also the general memorability of scene pictures. In addition, experiences with one's own learning and memory support the ability to assess scene memorability in general. This research contributes to our understanding of the basis and accuracy of different metamemory judgements.

Keywords

Metamemory accuracy; judgements of learning; memorability judgements; scene memorability

Received: 28 April 2023; revised: 30 October 2023; accepted: 26 February 2024

It is helpful to know which pictures are memorable. For instance, an illustrator may benefit from such knowledge when choosing pictures for advertisements. The ability to assess and to know about memory is termed as *metamemory* (Dunlosky & Thiede, 2013). An advantage of accurate metamemory is the effective regulation of future memory performance (Bjork et al., 2013). However, studies investigating metamemory accuracy of naturalistic scene pictures have yielded conflicting evidence: It is not clear how accurate people are at predicting which pictures will be remembered and which will not. Thus, although it is well known that the human visual memory storage for pictures is astonishing (Nickerson, 1965; Shepard, 1967; Standing, 1973), how good metamemory for pictures is remains to be examined.

The conflicting findings on metamemory accuracy for scene pictures stem from studies using either *memorability judgements* (MJs)—judgements of stimulus memorability in general—or *judgements of learning* (JOLs)—predictions of one's own later memory performance for recently studied items. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) found that MJs are unpredictive of actual picture memorability. This is surprising since different people tend to remember and forget the same pictures (Isola, Parikh, et al., 2011; Isola,

[1] Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany
[2] Department of Psychology, Technical University of Darmstadt, Darmstadt, Germany

Corresponding author:
Sofia Navarro-Báez, Alexanderstr. 10 (S1|15), 64283 Darmstadt, Germany.
Email: sofia.navarro@tu-darmstadt.de

Xiao, et al., 2011, 2014). In contrast, JOLs have been found to be moderately predictive of actual individual recognition memory for pictures (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). The current study aims to test whether these differences are due to MJs referring to memorability as a generic item attribute, whereas JOLs refer to one's own chances of remembering a recently studied item. This endeavour will enhance our understanding of the accuracy and basis of MJs and JOLs and extend our knowledge about different metamemory judgements.

Memorability of scene pictures

Although items may naturally vary in their actual memorability between individuals due to idiosyncratic encoding (Hintzman, 1980; Undorf et al., 2022), recent work has indicated that actual memorability of scene pictures is quite consistent across participants (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011, 2014). In a large-scale series of studies, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) measured the memorability of more than 2,000 images of real-world scenes from the SUN database (Xiao et al., 2010). They used a repeat detection task in which participants saw sequences of 120 images and were asked to detect whenever there was a repetition of an image. *Image memorability* was measured as the percentage of correct detections by participants. To investigate how consistent image memorability is across participants, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) randomly split the sample into two independent halves and correlated the image memorability values from the two halves. Repeating this procedure over 25 times, the average correlation was strong ($\rho = .75$) and indicated that people tend to recognise and miss the same pictures.

Given consistency of image memorability across participants, a further step is to explain what makes an image memorable. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) identified attributes contributing to memorability. Highly memorable images had semantic attributes such as enclosed spaces, telling a story, and people present. In contrast, less memorable images displayed open spaces, aesthetic settings, and were peaceful. Interestingly, perceptual image features such as colour (e.g., hue, saturation) and object statistics (e.g., number of objects, coverage of pixels over objects) were unrelated to memorability. Overall, image memorability was mainly predicted by the high-level semantic information conveyed in the picture (but see Lin et al., 2021). Nevertheless, a large proportion of image memorability variance remained unexplained.

MJs

If image memorability tends to be the same across participants, it is reasonable to ask whether people can assess the memorability of pictures. To address this question, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) obtained MJs in two tasks; in the first task, 30 participants were asked, "Is this a memorable image? Yes/No," and in the second task, 30 other participants were asked, "If you were to come across this image in the morning, and then happen to see it again at the end of the day, do you think you would realize that you have seen this image earlier in the day? Yes/No." Results showed that MJs did not predict image memorability: correlations between MJs and memorability were $\rho = -0.19$ in the first task and $\rho = -0.02$ in the second task. Instead, MJs were highly correlated with average ratings of semantic image attributes from a norming sample (aesthetics, $\rho = .83$; interestingness, $\rho = .86$) that were inversely related with image memorability (aesthetics, $\rho = -.36$; interestingness, $\rho = -.23$). These results suggest that people have the misconception that beautiful and interesting images are highly memorable and, more generally, indicate that people lack insight into item memorability.

JOLs

JOLs are commonly studied metamemory judgements. When people make JOLs, they predict their own future memory performance for recently studied items. Crucially, JOLs are elicited after learning each item and are compared with participant's own later memory performance. Higher-order monitoring processes of learning and memory are involved when making JOLs (Nelson & Narens, 1990). Inferential accounts of metamemory assume that JOLs are inferences based on available cues and heuristics because there is no direct access to the strength of the memory trace (Koriat, 1997). Cues for JOLs are classified into three different types (Koriat, 1997). *Intrinsic cues* are characteristics inherent to the studied items, such as word concreteness or the aesthetics of a picture. *Extrinsic cues* are related to the study conditions in which items are learned, such as the number of study repetitions and encoding strategies used. *Mnemonic cues* are sensitive to the effects of extrinsic and intrinsic cues and derive from the quality of processing items during learning, such as ease of encoding or retrieval fluency.

Evidence for inferential accounts of metamemory comes from situations in which metamemory judgements are dissociated from actual memory, leading to metamemory illusions (see Undorf, 2020; Undorf et al., 2022, for a review). For pictorial materials, there have been very few illusions found. One of them is for picture emotionality. Recognition memory performance is reduced for emotional pictures, but JOLs tend to be higher for emotional pictures compared with neutral ones (Caplan et al., 2019; Hourihan, 2020; Hourihan & Bursey, 2017). However, on a free recall test in which participants verbally described studied pictures, JOLs accurately predict better memory for emotional pictures (Schmoeger et al., 2020; Tauber et al., 2017). Thus, people recognise the positive validity of picture emotionality on free recall but fail to consider the differential negative effects of emotionality in a recognition memory test.

Accurate JOLs, in contrast, imply that these are based on cues that are predictive of people's actual memory performance (Chandler, 1994; Dunlosky & Metcalfe, 2009; Koriat, 1997). An important aspect of metamemory accuracy is relative accuracy (or resolution)—the extent to which metamemory judgements discriminate between items that will be remembered and those that will not be remembered. Importantly, most of the few JOL studies using pictures of scenes have found that JOLs are relatively accurate in terms of relative accuracy and track cue effects on actual memory performance (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This is illustrated in the study by Undorf and Bröder (2021), in which a total of six intrinsic and extrinsic cues in pictures from the SUN database (Xiao et al., 2010) were manipulated across three experiments. Results showed that recognition memory performance was better for scenes that were contextually distinctive, coloured (vs. grayscale), telling a story, twice (vs. once) presented, and containing persons, whereas recognition memory performance was worse for peaceful scenes. At the same time, people's JOLs were higher for all cues that helped memory and only failed to reflect that peacefulness hindered memory. Moreover, JOLs showed moderate relative accuracy, suggesting that reliance on valid probabilistic cues is an important factor for relative accuracy. Similarly, studies with verbal materials found that JOLs are based on multiple cues most of which have predictive validity and are moderate in their relative accuracy (e.g., Bröder & Undorf, 2019; Koriat, 1997; Undorf et al., 2018).

Differences between JOLs and MJs

As mentioned above, prior research by Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) suggest that MJs are unpredictive of actual image memorability across participants, while JOLs are relatively accurate at predicting participants' own memory performance (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This is interesting because both judgements refer to picture memorability and, consequently the same judgement target.

A potential reason for differences in accuracy between JOLs and MJs could be that they refer to different aspects of memorability. JOLs are predictions of one's own memory performance, whereas MJs are estimations of memorability as a general attribute (i.e., picture memorability) and do not focus on one's own experiences during learning and remembering. It is possible that people use different cues to inform judgements about memorability as a general attribute as opposed to their own learning and memory (Tullis & Fraundorf, 2017). In addition, people may use different cues for metamemory judgements made during a learning task versus a judgement-only task. For instance, pre-study JOLs made prior to learning items based on information about cue levels only (e.g., "You are about to study an emotional item") show lower relative accuracy than standard immediate JOLs. This is because pre-study JOLs can only be based on beliefs about how cue values affect memorability (e.g., "It is an emotional item, so it is easy to remember"; Price & Harrison, 2017; Undorf & Bröder, 2020), but not on learning. Similarly, ease-of-learning judgements made prior to learning (e.g., "How easy or difficult will it be to learn this item?") show lower relative accuracy than immediate JOLs (Kelemen et al., 2000; Leonesio & Nelson, 1990; Pieger et al., 2016). Furthermore, JOLs that are elicited immediately after studying each item are less accurate than JOLs elicited with a delay (Dunlosky & Nelson, 1992, 1994; Nelson & Dunlosky, 1991). This is because the cues available after learning might be more diagnostic than those during learning where item information is still present in working memory. Taken together, JOLs might rely more on diagnostic cues than MJs because they are made for one's own memory during a learning task.

In addition, the accuracy between JOLs and MJs might differ because of the memory tasks used to measure picture memorability and the memory criterion value used for accuracy. Picture memorability measures may, for example, vary between memory tasks. JOL studies with pictures of scenes often use a learning phase followed by an old/new recognition memory test (Caplan et al., 2019; Hourihan, 2020; Hourihan & Bursey, 2017; Kao et al., 2005; Undorf & Bröder, 2021). In contrast, in their study on MJs, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) employed a repeat detection task in which participants simultaneously encoded images and detected image repetitions. Moreover, regarding the memory criterion value, JOLs are related to participant's own individual memory performance (i.e., correlation of JOLs with item recognition memory by participant), whereas MJs are related to image memorability at the item level (i.e., correlation of MJs with item recognition memory aggregated across participants from other samples). Although it has been demonstrated that image memorability is highly consistent across participants (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011, 2014), there might be individual differences contributing to judgement predictive accuracy. By aggregating recognition memory performance across participants, idiosyncratic information influencing memory and metamemory is not considered (Tullis & Fraundorf, 2017; Undorf et al., 2022). This might be another reason contributing to the lower accuracy of MJs.

It is important to mention that a recent study showed that judgements of perceived memorability and JOLs for pictures of real-world objects and faces were both predictive of actual stimulus memorability (Saito et al., 2023). Given differences in the stimuli materials, parallels between that research and the current study are difficult to draw. Scene pictures are complex and high-dimensional stimuli that cannot be easily recoded with a simple verbal label. In contrast, real-world objects are easier to encode and retrieve because they benefit from an imaginal/verbal dual-coding processing (Paivio, 1991). Also, face processing is highly specialised and may not be comparable to the processing of other visual stimuli (Bruce & Young, 1986; Schwaninger et al., 2004). Thus, MJs might have shown low accuracy in prior studies due to the complexity and high dimensionality of scene pictures.

The current study

The aim of this study was to directly compare the relative accuracy and cue basis of JOLs and MJs for pictures of scenes. To achieve this aim, participants made two types of metamemory judgements, JOLs and MJs, for different aspects of picture memorability (one's own future memory vs. memorability as a generic item attribute) and during a learning versus judgement-only task, respectively. At the same time, we ensured that the MJ and JOL procedures were as similar as possible in all other respects. Specifically, we used the same judgement scale for both judgements, manipulated identical cues, and investigated the relative accuracy of JOLs and MJs with respect to the same memory criterion, namely, actual population memorability of scenes. *Population scene memorability* was defined as the proportion of recognition hits minus the proportion of false alarms per scene in each experiment's recognition memory task. This measure corresponds to the proportion of corrected hit rates (also known as *Pr*, Snodgrass & Corwin, 1988) and prevents that false memories contribute to the actual memorability of scenes.[1] We also investigated the relative accuracy of JOLs with respect to the participants' own memory performance criterion.[2] If discrepant findings regarding the accuracy of JOLs and MJs reported in prior work are mainly due to differences in the judgements' cue basis, we expect to see clear differences in cue use and judgement accuracy across JOLs and MJs. In contrast, if discrepant findings are largely due to methodological differences across studies obtaining JOLs and MJs, we expect to see similar accuracy and cue basis.

In Experiments 1 and 2, we used a within-subjects design by presenting a JOL task and an MJ task to the same participants, with the order of tasks counterbalanced between participants. In the JOL task, participants studied and made JOLs for a set of pictures, completed a distraction task, and finally completed a recognition memory test. In the MJ task, participants judged the memorability of another set of pictures. In Experiment 1, we orthogonally manipulated aesthetics and interestingness in two clearly distinguishable levels to compare the cue basis of MJs and JOLs. In Experiments 2 and 3, we used pictures that represented the whole range of normed image memorability in a continuous way. To foreshadow the results, we found that MJs and JOLs had similar cue bases and were both predictive of scene memorability, but that the relative accuracy of MJs improved after a JOL task (Experiment 1). This effect was completely unexpected, so we replicated it in Experiment 2. To gain further insight in this unexpected and theoretically relevant result, we designed Experiment 3 to disentangle which component of the JOL task drives the improvement in MJ accuracy. For this, we used a four-group design in which participants completed either (1) the learning phase with JOLs and a memory test (i.e., the full JOL task as in the previous experiments), (2) the learning phase without JOLs plus a memory test, (3) the learning phase with JOLs but no memory test, or (4) no learning phase with JOLs and no memory test (i.e., no component of the JOL task) before completing the MJ task. We found that the learning phase by itself was sufficient for the improvement in MJ accuracy. In addition, we found that MJs were more sensitive to cue effects after a memory test than after making JOLs. This was in line with Pearson correlations showing higher MJ accuracy when participants previously took a test than when they made JOLs.

Experiment 1

In Experiment 1, we examined the relative accuracy and cue basis of JOLs and MJs for pictures of naturalistic scenes that varied in aesthetics and interestingness. Aesthetics and interestingness were the image attributes identified as negative predictors of image memorability, but positively affecting MJs in Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)). The JOL task was similar to the one used by Undorf and Bröder (2021). Participants provided a JOL after studying each picture from a set of 120 pictures, and, following the learning phase, completed a recognition memory test with 240 pictures. In the MJ task, participants gave an MJ for each picture from another set of 120 pictures. They were explicitly instructed not to study the pictures, but only to judge their general memorability. This procedure was very similar to that used by Isola, Parikh, et al. (2011), Isola, Xiao, et al. (2011, 2014), with the exception that Isola et al. obtained binary ratings, whereas we used the same 11-point scale for MJs and JOLs. This was critical to prevent that potential accuracy differences between the two types of judgements could stem from using different judgement scales. To manipulate aesthetics and interestingness, we presented scenes from all possible combinations of high and low interestingness and aesthetics to participants. As Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) found that aesthetics and interestingness negatively affected memory performance, we expected that memory performance for pictures would be worse for scenes high

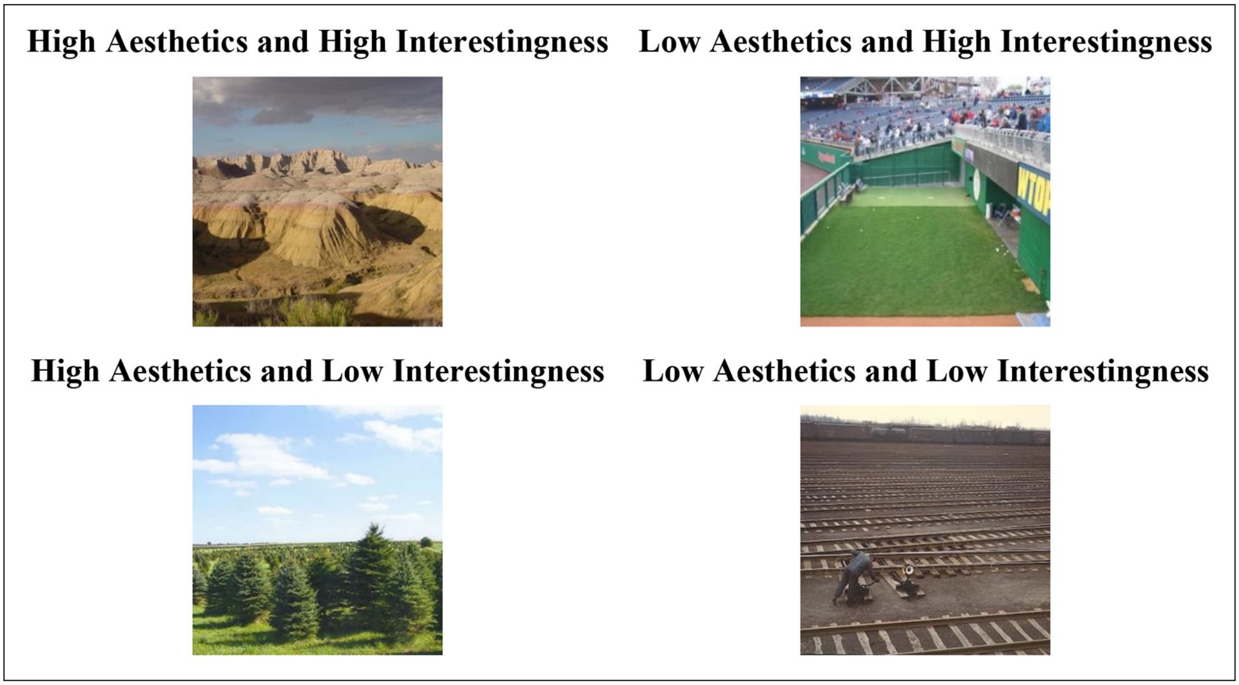
High Aesthetics and High Interestingness
Low Aesthetics and High Interestingness
High Aesthetics and Low Interestingness
Low Aesthetics and Low Interestingness

Figure 1. Example pictures for each combination of aesthetics and interestingness used in Experiment 1.

in aesthetics and interestingness. Furthermore, given that JOLs for pictures were usually moderately accurate and relied on valid cues (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021), we predicted accurate JOLs that would decrease with aesthetics and interestingness. It was an open question whether MJs would be accurate and relying on valid cues.

Method

Design and materials. The design was a 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed design, with aesthetics and interestingness as within-participants factors and task order as a between-participants factor. Half of the participants were randomly allocated to the JOLs-first condition ($n=26$). The other half of participants were allocated to the MJs-first condition ($n=26$). Aesthetics and interestingness were manipulated by selecting different sets of normed scene pictures. Stimuli were 360 pictures from the SUN database (Xiao et al., 2010). Normed values for aesthetics and interestingness were taken from Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) who asked 30 participants "Is this an aesthetic image?" and "Is this an interesting image?" Yes/No. Ninety scenes each were low in aesthetics and interestingness, low in aesthetics and high in interestingness, high in aesthetics and low in interestingness, and high in aesthetics and interestingness (see Figure 1).[3] We divided the scenes into three parallel sets each with 120 scenes in total, 30 of them from each combination of aesthetics and interestingness. For each participant, scenes from one randomly selected set served as study items in the JOL task, scenes from another randomly selected set served as distractors in the test phase of the JOL task, and scenes from the third set were used in the MJ task.

Participants. We aimed at recruiting at least 50 participants from the Prolific online subject pool. This sample size provides a statistical power of $(1-\beta)=.94$ to detect medium-sized main and interaction effects ($f=.25$, equivalent to $\eta_p^2=.06$) with $\alpha=.05$ in a mixed ANOVA when assuming a correlation of .50 between repeated measures (G*Power 3; Faul et al., 2007). We recruited participants who were 18–61 years old, reported English as their first language, and had at least a high school diploma as highest degree. The experiment took approximately 40 min and participants were paid £5.

To ensure high data quality, our criteria for not accepting submissions of participants in Prolific were: (1) study timed out, based on a time limit set by Prolific based on the estimated completion time ($n=4$), (2) completing the study with a different device than a desktop computer ($n=0$), or (3) low effort throughout the experiment operationally defined as writing gibberish in the filler task ($n=0$) or corrected hit rates of or very close to zero ($n=1$).[4] We accepted submissions from 57 participants. Our criteria for excluding accepted submissions from analysis were: participants reported technical problems ($n=5$), admitted having used helping tools during the study ($n=0$), or admitted having completed the study with the help of someone else ($n=0$). The final sample included 52 participants (37 females, 14 males, and 1 other). Their mean age was 32.42 years ($SD=11.1$), 3 participants were between 18 and 20 years in age, 29 participants were between 21 and 30 years in age, 7 participants were between 31 and 40 years in age, 7 participants were between 41 and 50 years in age, and 6 participants were between 51 and 61 years in age.

Procedure. The experiment consisted of a JOL task and an MJ task. Participants in the JOLs-first condition completed the JOL task first and then completed the MJ task. Task order was reversed for participants in the MJs-first condition. At the beginning of the experiment, we asked participants to comply with the following requirements: maximising the size of the web browser so that it covers the entire screen, completing the study in a single session, not leaving the study to engage in other tasks, completing the study in an environment that is free of noise or distraction, and not using any helping tools to complete the tasks.

In the JOL task, participants were instructed that their task was to remember 120 scene pictures for a later memory test in which studied photos would be intermixed with new ones and they would be asked to indicate whether each photo presented was studied or new. They were also instructed to predict the chances that they would personally recognise the photo on the test immediately after learning each photo. At learning, each scene picture was centred in the top half of the screen and displayed for 1 s, preceded by a 500-ms fixation cross that appeared in the same location. Immediately afterwards, participants indicated their chances of recognising the picture at test. To make their self-paced JOL, participants clicked on one of 11 buttons labelled 0, 10, . . ., 90, and 100. Following the learning phase, participants performed a semantic filler task for 3 min. On each filler trial, participants had 20 s to type in one word from each of three categories (i.e., animal, meal, and city) that started with a given letter. Finally, participants completed a self-paced recognition test with 240 scenes that included the 120 studied and 120 new scenes. At test, each scene picture was centred in the top half of the screen and participants indicated whether they had studied the picture before by clicking on buttons labelled "yes" and "no."

In the MJ task, participants were told that they would be presented with 120 scene pictures and their task is to judge how memorable each scene is. Participants were informed that they need not study the pictures themselves. At judging, each scene picture was centred in the top half of the screen and participants indicated how memorable the picture was for people who are asked to memorise the photo and later recognise it among new pictures. To make their self-paced MJ, participants clicked on one of 11 buttons labelled 0 (not memorable at all), 10, . . ., 90, and 100 (very memorable). For each participant, scene pictures were presented in a new random order in the learning phase and recognition memory test of the JOL task, and in the MJ task.

Data analysis

We report three different measures of judgement resolution. In all measures of judgement resolution, we used population scene memorability at the item level as memory criterion for MJ and JOL accuracy. Population scene memorability corresponds to the corrected hit rate for each scene, and it was calculated by subtracting the false alarm rate from the hit rate per scene. As the memory criterion for JOL accuracy at the individual level (i.e., participants' own memory performance), we used uncorrected hit rates because it is impossible to correct hit rates for both individual participants and individual items. Our main measure of judgement resolution is the within-subject Goodman–Kruskal gamma correlation between metamemory judgements and memory performance (Nelson, 1984). This is one of the most used measures of relative metamemory accuracy. Because population scene memorability was a continuous variable, we also report Pearson correlation coefficients. Furthermore, because gamma has been criticised due to inflated Type 1 errors (Murayama et al., 2014), discarded ties (Masson & Rotello, 2009; Spellman et al., 2014), and variation with liberal or conservative response criteria in recognition memory (Masson & Rotello, 2009), we additionally conducted a mixed-effects model analysis predicting population scene memorability from MJs and JOLs (Murayama et al., 2014).

Results

Resolution of JOLs and MJs. Table 1 and Figure 2 show mean gamma correlations between metamemory judgements and population scene memorability for each task in each task order condition. All correlations were significantly positive, $t>=5.02$, $p<.001$, indicating that not only JOLs but also MJs captured differences in population scene memorability. A 2 (task: JOL vs. MJ; within-participants) × 2 (task order condition: JOLs-first vs. MJs-first; between-participants) mixed ANOVA revealed no main effects, task: $F(1, 50)=2.24$, $p=.14$, $\eta_p^2=.04$, task order condition: $F(1, 50)=1.34$, $p=.25$, $\eta_p^2=.03$, but a significant interaction, $F(1, 50)=25.30$, $p<.001$, $\eta_p^2=.34$. Follow-up t-tests indicated that gamma correlations for JOLs did not differ between conditions, $t(50)=1.36$, $p=.18$, $d=0.38$, whereas gamma correlations for MJs were higher in the JOLs-first condition than in the MJs-first condition, $t(50)=3.09$, $p<.01$, $d=0.88$, which indicates higher relative accuracy of MJs when made after the JOL task. Equivalent results were found with the mixed-effects model analysis (see the Supplementary Material 2). Similar results were found with Pearson correlations, except for a main effect of task indicating higher Pearson correlations for MJs than for JOLs (see the Supplementary Material 1).

Table 1 shows mean gamma correlations between JOLs and participant's own memory performance (individual

Table 1. Means (*SDs*) of the gamma correlation between population scene memorability (hit rate corrected per scene) or participant's own memory performance and JOLs or MJs in each task order condition of Experiments 1, 2, and 3.

Experiment and condition	Accuracy criterion		
	Population scene memorability		Own memory performance
	JOLs	MJs	JOLs
Experiment 1			
JOLs first	.21 (.14)	.35 (.17)	.36 (.15)
MJs first	.27 (.17)	.19 (.19)	.40 (.16)
Experiment 2			
JOLs first	.23 (.14)	.31 (.15)	.33 (.25)
MJs first	.25 (.18)	.19 (.17)	.45 (.18)
Experiment 3			
MJ-task-only	-	.20 (.22)	-
Full-JOL-task	.26 (.14)	.32 (.15)	.38 (.22)
Study-and-JOL-task	.28 (.18)	.26 (.20)	-
Study-and-test-task	-	.31 (.17)	-

Note. JOLs = judgements of learning, MJs = memorability judgements, Population scene memorability = hit rate minus false alarm rate per scene across participants in each experiment.

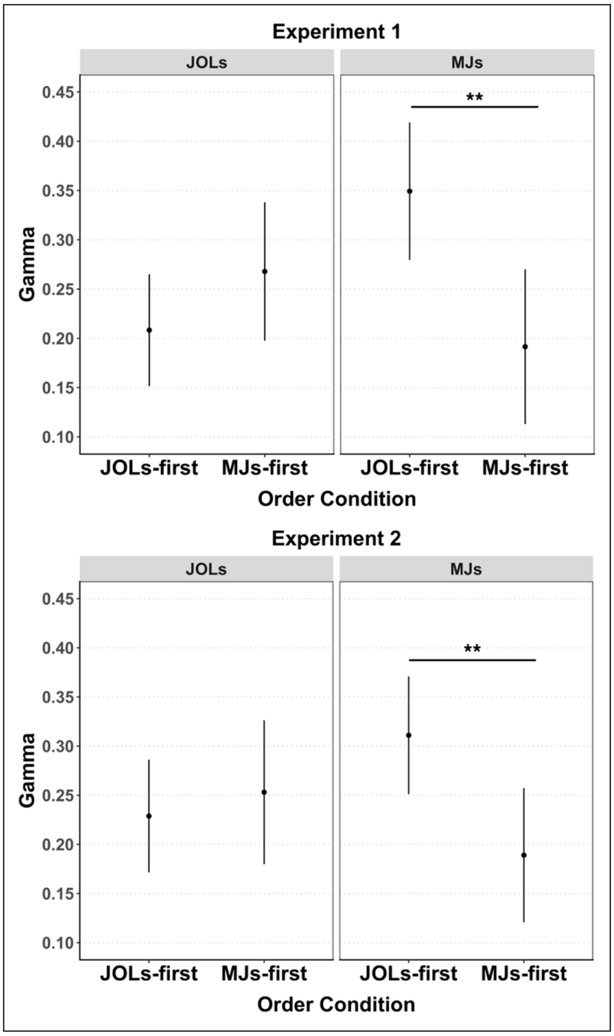

Figure 2. Gamma correlations between population scene memorability (hit rate corrected per scene) and judgements of learning (JOLs) or memorability judgements (MJs) in each task order condition of Experiments 1 and 2.
Note. Error bars represent one standard error of the mean.
Experiment 1: p = .003
Experiment 2: p = .008

memory performance) in each task order condition. Both gamma correlations were significantly positive, $t \geq 12.04$, $p < .001$, and they did not differ between order conditions, $t < 1$.

Cue effects on JOLs and individual memory performance. Figure 3 presents JOLs and corrected hit rates from the JOL task by aesthetics and interestingness in the JOLs-first and MJs-first condition. A 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed ANOVA on JOLs revealed no main effect of aesthetics, $F(1, 50) = 0.77$, $p = .$ 39, $\eta_p^2 = .02$, a main effect of interestingness, $F(1, 50) = 90.02$, $p < .001$, $\eta_p^2 = .64$, indicating higher JOLs for scenes high in interestingness than low in interestingness, no main effect of task order condition, $F(1, 50) = 1.64$, $p = .21$, $\eta_p^2 = .03$, and a significant interaction between order condition and interestingness, $F(1, 50) = 7.75$, $p < .01$, $\eta_p^2 = .13$. Follow-up *t*-tests indicated that interestingness affected JOLs in both conditions, but more so in the MJs-first condition, JOLs-first condition: $t(25) = 5.34$, $p < .001$, $d = 1.07$; MJs-first condition: $t(25) = 7.87$, $p < .001$, $d = 1.58$. No other interactions were significant, $F < 1$, $p \geq .37$.

A similar ANOVA on corrected hit rates (*Pr*) revealed better recognition memory performance for scenes low in aesthetics than high in aesthetics, $F(1, 50) = 54.73$, $p < .001$, $\eta_p^2 = .52$, and for scenes high in interestingness than low in interestingness, $F(1, 50) = 4.15$, $p = .047$, $\eta_p^2 = .08$, no other effects were significant, $F \leq 1.64$, $p \geq .21$.[5] We thus replicated Isola, Parikh, et al.'s (2011) and Isola, Xiao, et al.'s (2011, 2014) findings of better memory performance for scenes low in aesthetics, but did not replicate better memory performance for scenes low in interestingness. Instead, we found that memory performance was better for scenes high in interestingness. We will return to this point in the "Discussion" section.

Cue effects on MJs. Figure 3 presents MJs from the MJ task by aesthetics and interestingness in the JOLs-first and MJs-first condition. A 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed ANOVA on MJs revealed no main effect of aesthetics, $F(1, 50) = 3.65$, $p = .06$, $\eta_p^2 = .07$, a main effect of interestingness, $F(1, 50) = 224.02$, $p < .001$, $\eta_p^2 = .82$, indicating higher MJs for scenes high

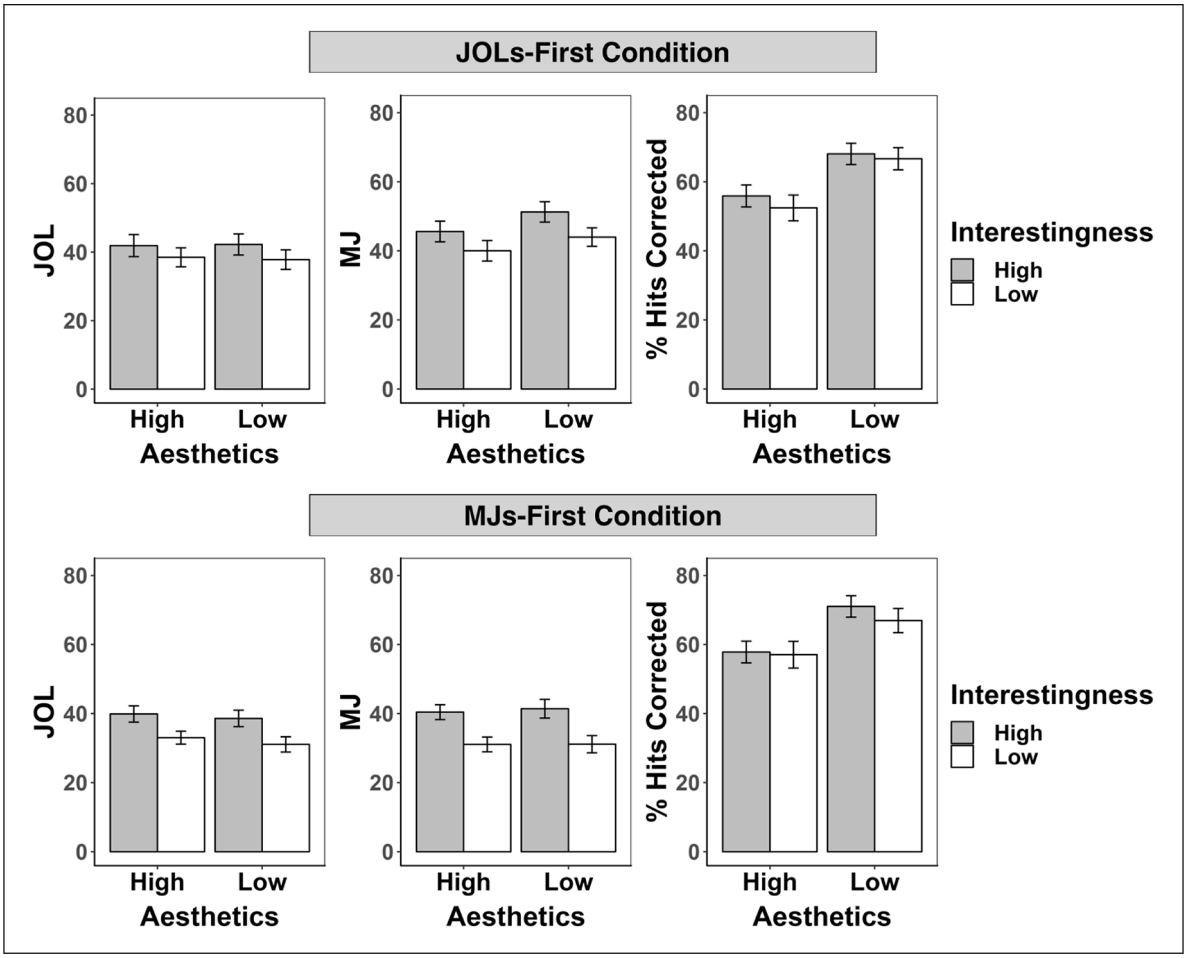

Figure 3. Mean judgements of learning (JOL), memorability judgements (MJ) and corrected hit rates (% hits corrected) by aesthetics and interestingness in the JOLs-first (top panel) and MJs-first (bottom panel) conditions of Experiment 1.
Note. Error bars represent one standard error of the mean.

in interestingness than low in interestingness, a main effect of task order condition, $F(1, 50) = 7.42$, $p < .01$, $\eta_p^2 = .13$, indicating higher MJs in JOLs-first condition than in the MJs-first condition, and a significant interaction between interestingness and task order condition, $F(1, 50) = 9.71$, $p < .01$, $\eta_p^2 = .16$. Follow-up *t*-tests indicated that interestingness affected MJs in both conditions, but more so in the MJs-first condition, JOLs-first condition: $t(25) = 9.33$, $p < .001$, $d = 1.87$; MJs-first condition: $t(25) = 11.71$, $p < .001$, $d = 2.34$. All other interactions, $F \leq 2.33$, $p \geq .13$.

Discussion

Recognition memory was affected by the two image characteristics aesthetics and interestingness. As in Isola et al., we found better memory performance for scenes low rather than high in aesthetics. Contrary to Isola et al., we found better memory performance for scenes high rather than low in interestingness. Potential explanations for this difference in results include that we used a different recognition memory paradigm (i.e., learning phase followed by an old/new memory test), and that aesthetics and interestingness were strongly correlated in Isola et al.'s study ($\rho = .85$) but manipulated orthogonally here. Furthermore, the finding that JOLs and MJs were both unaffected by aesthetics, but higher for pictures high rather than low in interestingness suggests a similar cue basis of JOLs and MJs. The finding that aesthetics did not affect either metamemory judgement fits with prior findings indicating that people sometimes fail to factor valid cues in their JOLs for scene pictures (e.g., peacefulness, Undorf & Bröder, 2021).

Despite people's failure to recognise the predictive validity of aesthetics in their JOLs and MJs, reliable resolution showed that both metamemory judgements captured differences in the relative population memorability of scenes. Thus, by directly comparing JOLs and MJs in a within-subjects design using the same memory criterion, our results showed that both types of judgement had moderate resolution.

A new and unexpected finding was that the accuracy of MJs improved substantially after completing a JOL task, whereas completing an MJ task did not affect JOL accuracy.[6] The order of the tasks also did not affect JOL accuracy when using participant's own memory performance as memory criterion. This finding shows that the accuracy of JOLs and MJs differs in whether it is affected by the order of the tasks. Previous metamemory studies using multiple study-test cycles for the same materials have reported changes in the resolution and absolute accuracy of JOLs. From the second study-test cycle onward, reliance on past memory performance increases JOL resolution and, at the same time, under-confidence lowers absolute accuracy of JOLs (Finn & Metcalfe, 2008; King et al., 1980; Koriat et al., 2006). Importantly, the current finding that MJ resolution improves after a JOL task is novel in that it demonstrates increased accuracy of judging the general memorability of *new* pictures after actively engaging in a learning phase with JOLs and a memory test.

In conclusion, the results of Experiment 1 indicate that (1) JOLs and MJs for scenes were predictive of population scene memorability and that (2) both types of metamemory judgements had a similar cue basis (i.e., based on interestingness, but not on aesthetics). A surprising finding was that (3) having completed a learning task with JOLs and a recognition memory test improved the accuracy of MJs, whereas making MJs did not improve the accuracy of JOLs. A potential mechanism for this improvement in relative accuracy is that participants gained experience by intentionally learning pictures and reflecting about their own memory performance. If MJs are made without this experience, participants probably lack knowledge about how to assess the abstract image feature memorability. This interpretation suggests that experiences with one's own memory precede the understanding of memory in general, and it would help to explain why MJ and JOL accuracy is sometimes comparable and sometimes not. To rule out that improved MJ accuracy after the JOL task was an accidental result in Experiment 1, Experiment 2 aimed to conceptually replicate this finding.

Experiment 2

Experiment 2 aimed to replicate findings of Experiment 1 and, specifically, the unexpected finding that having completed a JOL task with learning pictures, making JOLs, and taking a recognition memory test improved the relative accuracy of MJs. Because JOLs and MJs were based on similar cues in Experiment 1, we did not manipulate individual cues in Experiment 2, but instead used scenes that varied widely in scene memorability. Based on the findings obtained in Experiment 1, we expected that both JOLs and MJs would be similarly impacted by scene memorability. As in Experiment 1, all participants completed a JOL task and an MJ task with the task order manipulated between participants. We expected that, as in Experiment 1, JOLs and MJs would be predictive of population scene memorability. At the same time, given the experience and knowledge people gained in the JOL task, we hypothesised that the relative accuracy of MJs would be higher in the JOL-first condition than in the MJ-first condition.

Method

Design and materials. The design was a 10 (scene memorability: 10 levels from low to high) × 2 (task order condition: JOLs-first vs. MJs-first) mixed design, with scene memorability as a within-participants factor and task order as a between-participants factor. Scene memorability was manipulated by selecting different sets of scene pictures that varied in corrected hit rates (i.e., hit rate minus false alarm rate per scene) reported in Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014). We used the deciles of the frequency distribution of scene memorability as cutoff values and selected 36 scenes from each of the 10 levels, resulting in a total of 360 pictures (see Figure 4).[7] We divided the scenes from each level of memorability into three parallel sets with 12 scenes. Recombining these, we thus created 3 parallel sets of 120 pictures each (12 of each level). Sets were also similar in aesthetics and interestingness. For each participant, scenes from one randomly selected set served as study items in the JOL task, scenes from another randomly selected set served as distractors in the test phase of the JOL task, and scenes from the third set were used in the MJ task.

Participants. We aimed at recruiting 50 participants from the Prolific online subject pool who were 18 to 61 years old, reported English as their first language, and had at least a high-school diploma as highest degree. Power analysis was identical to that of Experiment 1. The experiment took approximately 40 min and participants were paid £5. Based on the same criteria as in Experiment 1, we did not accept submissions in Prolific when the study timed out ($n=2$), was completed on a different device than a desktop computer ($n=1$), or there was low effort throughout the experiment ($n=0$).

We accepted submissions from 55 participants. Based also on the same criteria of Experiment 1, we excluded accepted submissions from analysis when participants reported technical problems ($n=5$), admitted having used helping tools during the study ($n=0$), or admitted having completed the study with the help of someone else ($n=0$). The final sample included 50 participants (30 females, 20 males). The mean age of participants was 34.64 ($SD=9.94$), 2 participants were between 18 and 20 years in age, 21 participants were between 21 and 30 years in age, 13 participants were between 31 and 40 years in age, 8 participants were between 41 and 50 years in age, and 6 participants were between 51 and 61 years in age.

Procedure. The procedure was identical to Experiment 1.

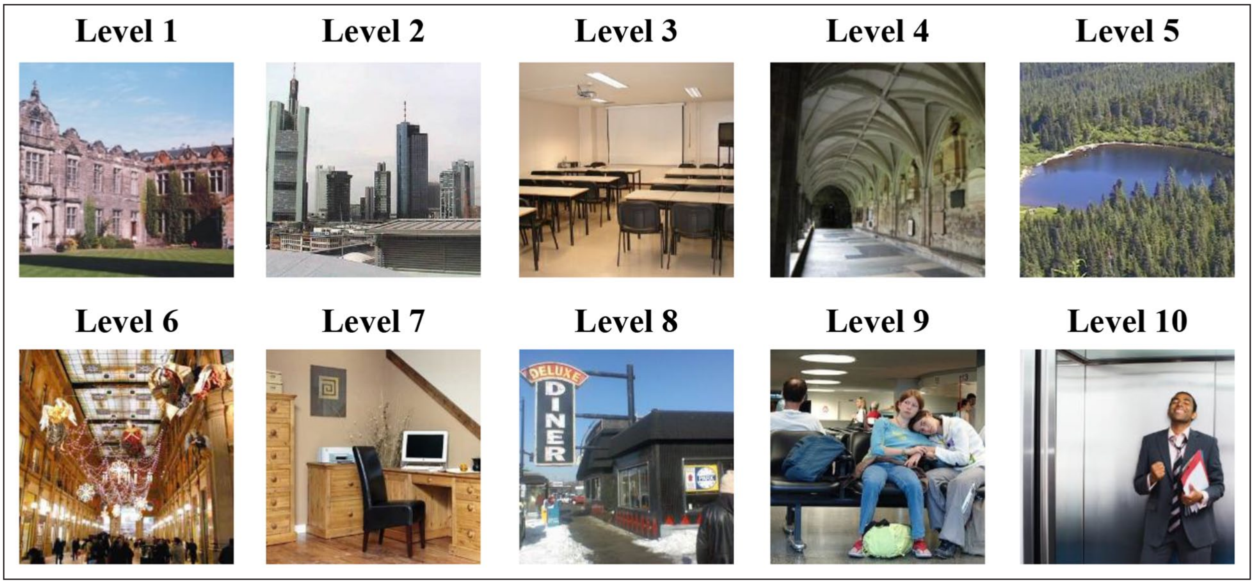

Figure 4. Example pictures of each of the 10 levels of scene memorability (from 1 = lowest to 10 = highest) used in Experiments 2 and 3.

Results

Resolution of JOLs and MJs. Table 1 and Figure 2 show mean gamma correlations between metamemory judgements and population scene memorability for each task in each task order condition. All correlations were significantly positive, $t > 5.72$, $p < .001$. A 2 (task: JOL vs. MJ; within-participants) × 2 (task order condition: JOLs-first vs. MJs-first; between participants) mixed ANOVA revealed no main effects: task, $F < 1$, task order condition, $F\ (1,\ 48) = 1.52$, $p = .22$, $\eta_p^2 = .02$, but a significant interaction, $F(1,\ 48) = 12.99$, $p < .001$, $\eta_p^2 = .21$. Planned comparisons indicated that gamma correlations for JOLs did not differ between conditions, $t < 1$, $p = .59$, whereas gamma correlations for MJs were higher in the JOLs-first condition than in the MJs-first condition, $t(48) = 2.77$, $p < .01$, $d = 0.80$. As in Experiment 1, this again shows higher relative accuracy of MJs after learning items, making JOLs, and completing a recognition memory test. Equivalent results were found with Pearson correlations and a mixed-effects model analysis (see the Supplementary Materials 1 and 2).

Table 1 shows mean gamma correlations between JOLs and participant's own memory performance in each task order condition. Both correlations were significantly positive, $t \geq 6.65$, $p < .001$, and did not differ between conditions, $t(48) = 1.88$, $p = .07$, $d = 0.57$.

Cue effects. Figure 5 presents JOLs, corrected hit rates, and MJs. We used a mixed-effects model (Bates et al., 2015) to evaluate whether JOLs and MJs increased monotonically with scene memorability. This approach allowed us to directly test for a linear increase in metamemory judgements with scene memorability as a fixed-effects predictor with 10 levels. To evaluate whether the scene memorability slope differed between order conditions, we included order condition and its interaction with scene memorability as additional fixed-effects predictors in the model. We specified random intercepts for participants and uncorrelated random slopes for scene memorability. Scene memorability was mean-centred, and task order condition was effect coded (−1 = MJs-first, 1 = JOLs-first). We used a logistic regression model to evaluate a linear increase in hit rates with scene memorability.

Cue effects on JOLs and individual memory performance. Regressing JOLs on scene memorability, task order condition, and their interaction revealed a significantly positive unstandardized coefficient for scene memorability, $b = 2.01$, ($SE = 0.22$), $t = 9.18$, $p < .001$, indicating that JOLs increased with scene memorability. No other effects were significant, order condition: $b = 4.02$, ($SE = 2.01$), $t = 2.00$, $p = .05$, interaction: $t < 1$. A logistic regression model revealed that hit rates increased with scene memorability, $b = 0.20$, ($SE = 0.01$), $z = 17.72$, $p < .001$. No other effects were significant, $z \leq 1.41$, $p \geq .16$.

Cue effects on MJs. Regressing MJs on scene memorability, task order condition, and their interaction revealed significantly positive unstandardized coefficients for scene memorability, $b = 1.91$, ($SE = 0.21$), $t = 8.90$, $p < .001$, indicating that MJs increased with scene memorability. The model also revealed significantly positive unstandardized coefficients for order condition, $b = 7.38$, ($SE = 1.79$), $t = 4.12$, $p < .001$, indicating that MJs were higher in the JOLs-first condition, and for the interaction between scene

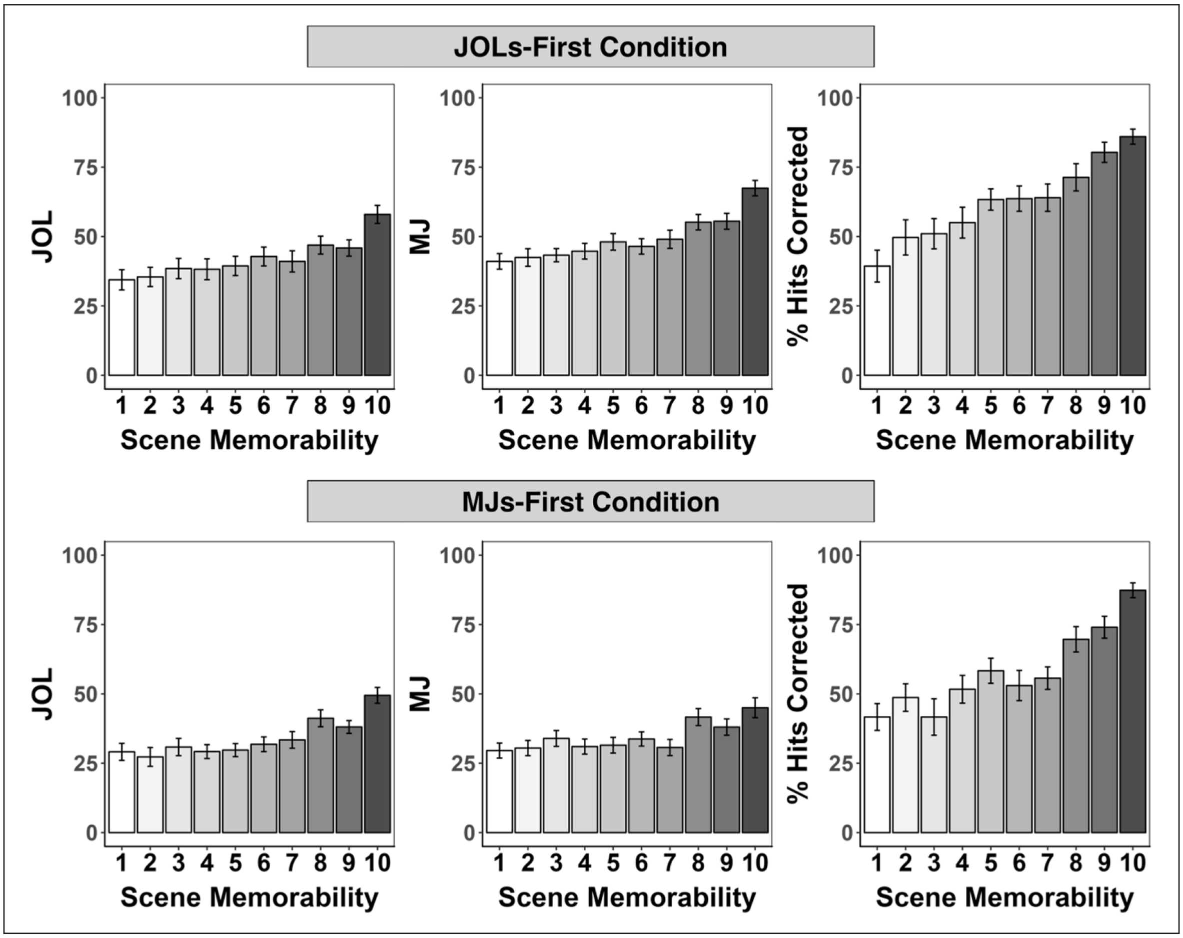

Figure 5. Mean judgements of learning (JOL), memorability judgements (MJ), and corrected hit rates (% hits corrected) by scene memorability in the JOLs-first (top panel) and MJs-first (bottom panel) conditions of Experiment 2.
Note. Error bars represent one standard error of the mean.

memorability and order condition, $b=0.51$, ($SE=0.21$), $t=2.38$, $p=.022$, indicating differences in the effects of scene memorability on MJs across order conditions. Separate follow-up regression models for each order condition revealed that MJs increased with scene memorability in both conditions, but more so in the JOLs-first condition, JOLs-first condition: $b=2.42$, ($SE=0.28$), $t=8.52$, $p<.001$; MJs-first condition: $b=1.40$, ($SE=0.32$), $t=4.36$, $p<.001$.

Discussion

As in Experiment 1, both JOLs and MJs were predictive of differences in population scene memorability. We again found that the relative accuracy of MJs improved after a JOL task, whereas prior experiences with making MJs did not improve JOL accuracy. Thus, when using the same memory criterion for accuracy at the item level, differences in accuracy between JOLs and MJs tied to the order of tasks still emerged. Importantly, this shows that the beneficial impact of a preceding JOL task on MJ accuracy is a robust effect that merits further scrutiny. The finding that task order did not affect JOL resolution was independent of the memory criterion used for accuracy (i.e., population scene memorability, or participant's own memory performance).

As expected, both JOLs and MJs monotonically increased with increasing scene memorability. This again indicated that the cue basis of the two metamemory judgements is similar and suggested that several cues diagnostic of memorability underlie each type of metamemory judgement (for evidence that multiple cues are integrated in JOLs for scene pictures and for verbal materials, see, for example, Undorf & Bröder, 2021; Undorf et al., 2018). Importantly, MJs increased more strongly with scene memorability in the JOLs-first than in the MJs-first condition, indicating that MJs become more sensitive to scene memorability effects after a JOL task. In contrast, scene memorability effects on JOLs were unaffected by the task order condition. This finding supports our hypothesis that participants learn about the general memorability of scenes by completing a JOL task and make MJs for new set of pictures on an updated basis.

In summary, results from Experiment 2 again show that experience with a JOL task provides a viable basis for assessing the general memorability of scenes. Therefore, this novel result from Experiment 1 was proven to be replicable and one may ask for an explanation. One step in this direction is to investigate which component of the JOL task (i.e., learning phase, making JOLs, recognition memory test) drives the improvement in MJ accuracy. Regarding the potential contribution of a learning phase to the relative accuracy of metamemory, literature is scarce. Two studies investigating the effects of prior learning versus prior testing on metamemory accuracy found that test experience was more effective than learning experience (Jang et al., 2012; Koriat & Bjork, 2006a). However, in the current study, it might be possible that a learning phase provides participants with the experience required to make accurate MJs. Regarding JOL experience, making JOLs for oneself might increase MJ accuracy because monitoring one's own learning can increase sensitivity towards diagnostic cues. For instance, previous studies have found that processing fluency as indicated by short self-paced study times is used as a cue for other's memory predictions only after learners had made JOLs for their own memory (Koriat & Ackerman, 2010; Undorf & Erdfelder, 2011). Therefore, from a cue-weighting perspective (Undorf et al., 2018), it may be that completing a learning phase with JOLs fosters the use of valid cues for MJs. These valid cues might include mnemonic cues such as the ease of encoding (Begg et al., 1989; Chandler, 1994; Hertzog et al., 2003) or perceiving pictures (Besken, 2016; Fei-Fei et al., 2007; Undorf et al., 2017) and intrinsic cues such as emotionality and concreteness (Undorf & Bröder, 2020). Alternatively, or additionally, test experience might improve MJ accuracy for a new set of scenes by providing participants with feedback regarding the memorability of scene pictures. Specifically, monitoring one's recognition memory performance for scene pictures during the test may provide hints of the features or feature combinations that make a picture memorable (Mitton & Fiacconi, 2020). For example, a participant might realise during the test that she recognises interesting pictures or pictures with people better than others. Thus, based on prior work, it is plausible that learning scene pictures, making JOLs, and taking a recognition memory test underlie the improvement in MJ accuracy observed in the previous experiments separately or in combination.

Experiment 3

Experiment 3 was preregistered (https://osf.io/3fujm) and aimed at disentangling which component of the JOL task drives the improvement in MJ accuracy. For this, in the first part of the experiment, different groups of participants completed either the full JOL task (*full-JOL-task* condition), a learning phase with JOLs, but without a test (*study-and-JOL-task* condition), a learning phase without JOLs, but with a recognition memory test (*study-and-test-task* condition), or no component of the JOL task (*MJ-task-only* condition). In the second part of the experiment, all participants completed an MJ task. In this four-group design, making JOLs (yes, no) and taking a test (yes, no) are fully crossed with the MJ-task-only condition being the control. However, the MJ-task-only condition (i.e., no JOLs, no test) additionally differs from the other three conditions by not including a learning phase. Because one cannot make JOLs or take a test without having learned the pictures, completing the learning phase (yes, no) cannot be fully crossed with the other variables (i.e., making JOLs, taking a test). Nevertheless, the imbalanced design allows for all crucial tests: If all experimental conditions show a similar improvement in MJ accuracy compared with the control condition, then completing a learning phase is the critical factor driving MJ accuracy. If MJ accuracy is higher in the full JOL-task condition than in the conditions in which participants make JOLs but do not take a test or take a test but do not make JOLs, then making JOLs and taking a test have additive effects on MJ accuracy. Finally, differences in MJ accuracy across the conditions in which participants make JOLs but do not take a test or take a test but do not make JOLs will reveal the relative importance of making JOLs or taking a test for improved MJ accuracy.

Method

Design and materials. The design was a 10 (scene memorability: 10 levels from low to high) × 4 (condition: full-JOL-task, study-and-test-task, study-and-JOL-task, MJ-task-only) mixed design, with scene memorability as a within-participants factor and condition as a between-participants factor. We used the same sets of pictures as in Experiment 2.

Participants. We aimed at recruiting $N=212$ participants from the Prolific online subject pool ($n=53$ in each condition) who were 18 to 61 years old, reported English as their first language, and had at least a high-school diploma as highest degree. This sample size provides a statistical power of $(1-\beta)=.95$ to detect medium-sized effects ($f=.25$, equivalent to $\eta_p^2=.06$) with $\alpha=.05$ in a fixed-effects ANOVA employed to test power for contrasts with $df=1$ and $df=4$ in the numerator and the denominator, respectively (G*Power 3; Faul et al., 2007; Perugini et al., 2018). The experiment took approximately 40 min and participants were paid £5. Participants were randomly allocated to one of the four conditions. Based on the same criteria as in Experiments 1 and 2, we did not accept submissions in Prolific when the study timed out ($n=2$), was completed on a different device than a desktop computer ($n=1$), or there was low effort throughout the experiment ($n=0$). We accepted submissions from 212 participants. Based also on the same criteria as Experiments 1 and 2, we excluded data from analysis when participants reported technical problems ($n=0$), admitted having used helping tools during the study ($n=3$), admitted completing the study with the help of someone else ($n=4$), or admitted having just clicked through the study without taking part seriously ($n=0$). The final sample included 205 participants ($n=51$ in the full-JOL-task, study-and-JOL-task, and MJ-task-only condition, $n=52$ in the study-and-test-task condition). They were 113 females, 91 males, and 1 other. The mean age of participants was 37.29 years ($SD=10.54$), 6 participants were between 18 and 20 years in age, 59 participants were between 21 and 30 years in age, 60 participants were between 31 and 40 years in age, 52 participants were between 41 and 50 years in age, and 28 participants were between 51 and 61 years in age.

Procedure. The experiment consisted of two parts. In the first part of the experiment, participants in the full-JOL-task condition completed the same JOL task as in Experiments 1 and 2 (i.e., learning phase with JOLs, semantic filler task, and recognition memory test). Participants in the study-and-JOL-task condition completed a learning phase with JOLs, and a semantic filler task, but no recognition memory test. They received the same initial instructions as participants in the full-JOL-task condition but learned at the end of the experiment that we wanted to examine the accuracy of memorability estimates in participants who had not taken a memory test, which is why they had skipped the memory test. Participants in the study-and-test-task condition completed a learning phase without JOLs, a semantic filler task, and a recognition memory test. Participants in the MJ-task-only condition completed the semantic filler task only. In the second part of the experiment, participants from all conditions completed the same MJ task as in Experiments 1 and 2.

Results

Resolution of JOLs and MJs. As in Experiments 1 and 2 and as preregistered, we examined the extent to which MJs predicted differences in the actual population memorability of scenes in this experiment using within-subject Goodman–Kruskal gamma correlations, Pearson correlations (see the Supplementary Material 1), and a mixed-effects model analysis (see the Supplementary Material 2). To assess differences in MJ accuracy between conditions, we transformed our hypotheses into a set of orthogonal contrasts using Helmert coding. This approach has two main advantages: (1) testing specific group differences that are independent and more informative than an omnibus F-test, and (2) greater statistical power than follow-up t-tests (Rosenthal et al., 2000). The first contrast tested the difference between the control condition (MJs-only; −3/4) and all three experimental conditions (full-JOL-task, study-and-JOL-task, study-and-test-task; coded all as +1/4). The second contrast tested the difference between the full-JOL-task condition (+2/3) and the other two experimental conditions (study-and-JOL-task group, the study-and-test-task group; coded both as −1/3). The third contrast tested the difference between the study-and-JOL-task condition (−1/2) and the study-and-test-task condition (+1/2).

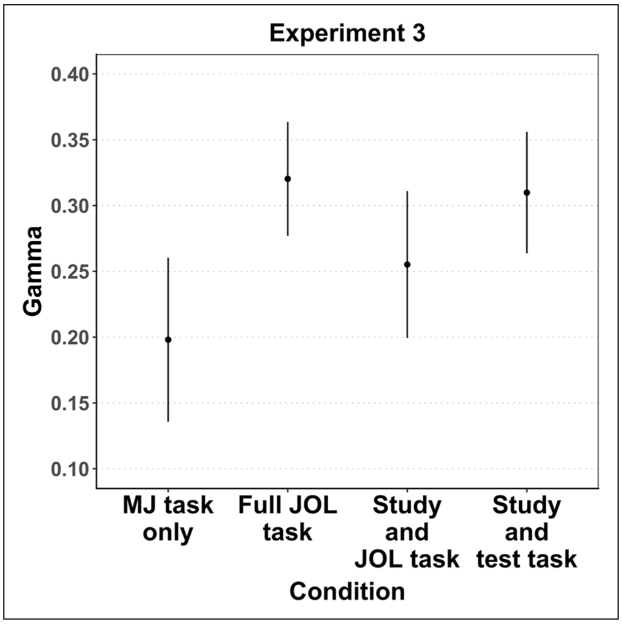

Figure 6. Gamma correlations between population scene memorability (hit rate corrected per scene) and memorability judgements (MJs) in each condition of Experiment 3.
Note. Error bars represent one standard error of the mean.

Table 1 and Figure 6 show mean gamma correlations between MJs and population scene memorability in each condition of Experiment 3. All correlations were significantly positive, $t \geq 6.39$, $p < .001$, indicating that MJs in all conditions captured differences in population scene memorability. Planned contrasts revealed that the learning phase present in all experimental conditions improved MJ accuracy compared with only making MJs, $t(201)=3.22$, $p < .01$. They also revealed that MJ accuracy did not differ between the full-JOL-task condition and the conditions with one component of the JOL task only (i.e., study-and-JOL-task, study-and-test-task), $t(201)=1.18$, $p=.24$, showing that there were no additive effects of making JOLs and taking a test. Finally, MJ accuracy did not differ between the study-and-JOL-task condition and the study-and-test-task condition, $t(201)=1.49$, $p=.14$, suggesting that making JOLs is not more beneficial than taking a test,

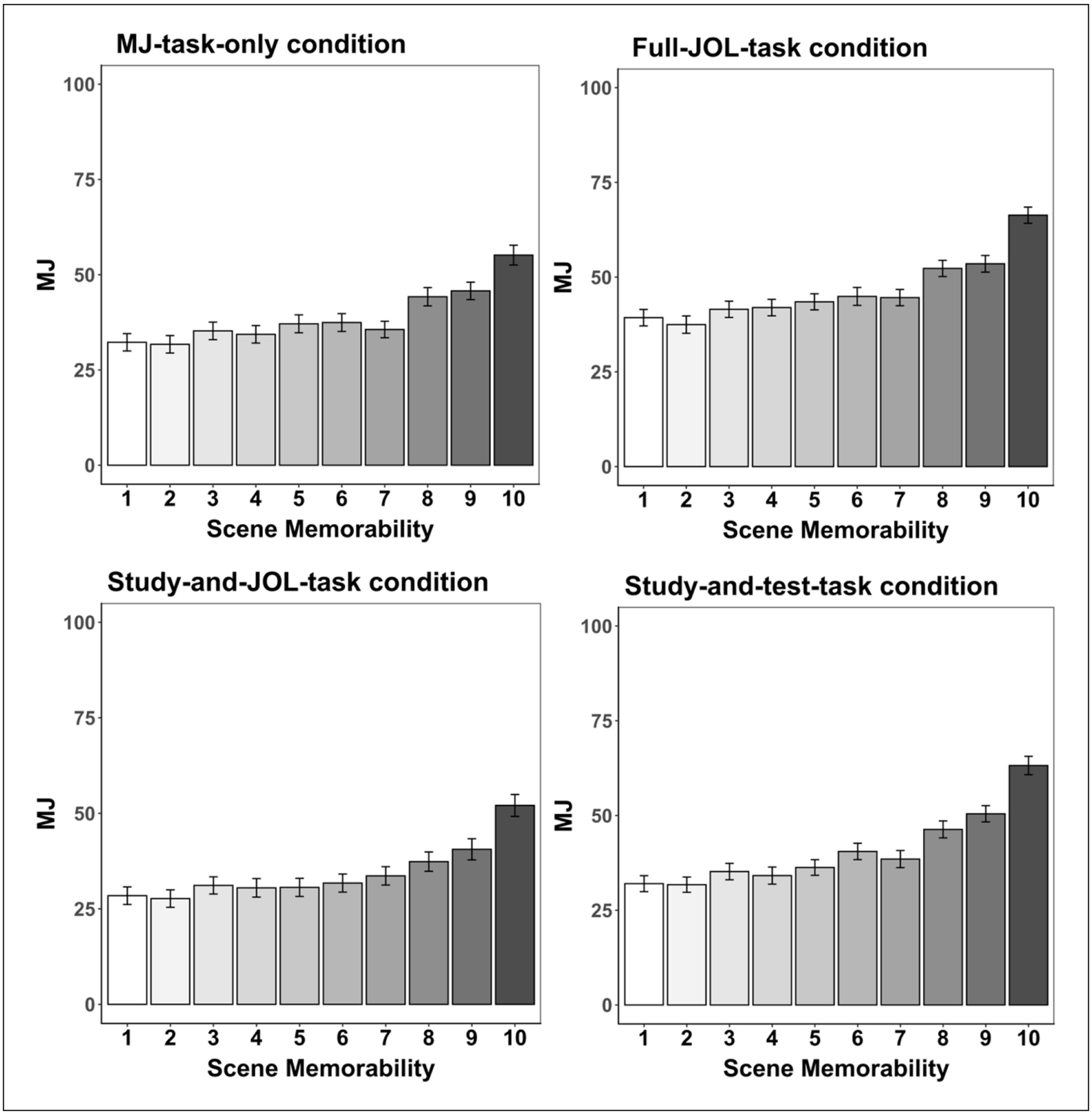

Figure 7. Mean memorability judgements (MJ) by scene memorability in each condition of Experiment 3.
Note. Error bars represent one standard error of the mean.

or vice versa. The latter result, however, was not supported by Pearson correlations which instead suggested that taking a test improved MJ accuracy more than making JOLs (see the Supplementary Material 1).

Cue effects on MJs. Figure 7 presents MJs in each condition of Experiment 3. As in Experiment 2 and as preregistered, we used a mixed-effects model to examine whether MJs increased with scene memorability. We included condition and its interaction with scene memorability as fixed-effects predictors in the model to evaluate whether the scene memorability slope differs between conditions. We specified random intercepts for participants and uncorrelated random slopes for scene memorability. Scene memorability was grand mean-centred, and condition was coded with the same Helmert contrasts as in the resolution analysis.

A significantly positive unstandardized coefficient for scene memorability, $b=2.43$, ($SE=0.19$), $t=22.43$, $p<.001$, indicated that MJs again increased with scene memorability. Significantly positive unstandardized coefficients for the second and third contrasts coding condition; $b=8.93$, ($SE=2.56$), $t=3.48$, $p<.001$, $b=6.47$, ($SE=2.95$), $t=2.19$, $p<.05$, indicated higher MJs in the full-JOL-task condition than in the study-and-JOL-task and the study-and-test-task conditions, and higher MJs in the study-and-test-task condition than in the study-and-JOL-task condition. More importantly, a significant interaction between scene memorability and the third contrast coding condition revealed differences in scene memorability effects on MJs between the study-and-JOL-task and the study-and-test-task conditions, $b=0.85$, $(SE=0.31)$, $t=2.78$, $p<.01$. Separate follow-up regression models for each condition revealed that MJs increased with scene memorability in both conditions, but more so in the study-and-test-task condition, study-and-JOL-task condition: $b=2.09$, $(SE=0.19)$, $t=11.05$, $p<.001$; study-and-test-task condition: $b=2.94$, $(SE=0.26)$, $t=11.51$, $p<.001$. All other effects were nonsignificant, $t\leq 1.52$.

Cue effects on JOLs and individual memory performance. We used a similar mixed-effects model to evaluate whether JOLs in the full-JOL-task and the study-and-JOL-task conditions increase with scene memorability. Condition was effect coded (−1 = study-and-JOL-task condition, 1 = full-JOL-task condition). This model revealed a significantly positive unstandardized coefficient for scene memorability, $b=2.17$, $(SE=0.13)$, $t=16.79$, $p<.001$, indicating that JOLs again increased with scene memorability. All other effects were nonsignificant, $t\leq 0.75$.

A logistic regression model was used to evaluate whether individual recognition memory performance in the full-JOL-task and the study-and-test-task conditions increases with scene memorability. Condition was effect coded (−1 = study-and-test-task condition, 1 = full-JOL-task condition). This model revealed that hit rates increased with scene memorability, $b=0.16$, $(SE=0.01)$, $z=20.40$, $p<.001$, that hit rates were higher in the full-JOL-task than in the study-and-test-task condition, $b=0.53$, $(SE=0.09)$, $z=5.88$, $p<.001$, and that scene memorability effects on hit rates differed between the full-JOL-task and the study-and-test-task conditions, $b=0.05$, $(SE=0.01)$, $z=5.85$, $p<.001$. Separate follow-up regression models for the latter two conditions revealed that hit rates increased with scene memorability in both conditions, but more so in the full-JOL-task condition, study-and-test-task condition: $b=0.11$, $(SE=0.01)$, $z=11.53$, $p<.001$; full-JOL-task condition: $b=0.20$, $(SE=0.01)$, $z=16.92$, $p<.001$.

Discussion

The aim of Experiment 3 was to disentangle which component of the JOL task drives the improvement in MJ accuracy observed in the previous experiments. All measures of resolution showed that MJ accuracy improved in all experimental conditions (full-JOL-task, study-and-JOL-task, study-and-test-task) in comparison to the control condition (MJs-only). As the learning phase is the common factor in all experimental conditions, our result suggests that a learning phase by itself provides experiences that are beneficial for subsequently assessing the memorability of pictures. Moreover, given that MJ accuracy was not better in the full-JOL-task condition than in the other two experimental conditions (study-and-JOLs-task condition, study-and-test-task condition), we did not find evidence for additive effects of making JOLs and taking a test on MJ accuracy. Regarding the individual effects of making JOLs and taking a test on MJ accuracy, gamma correlations suggested that neither making JOLs nor taking a test improves MJ accuracy, as did the mixed-effects model analysis. In contrast, the analysis of Pearson correlations reported in the Supplementary Material 1 suggested that a recognition memory test improves MJ accuracy more than making JOLs.

As in Experiment 2, we found that scene memorability influenced MJs, JOLs, and recognition memory performance. Importantly, MJs were influenced more strongly by scene memorability in the study-and-test-task condition than in the study-and-JOL-task condition. This result suggests that completing a recognition memory test is more beneficial for MJ accuracy than making JOLs, which is consistent with the Pearson correlation analysis showing that MJ accuracy is higher when having previously taken a test than having made JOLs, but not with the gamma correlation analysis or the mixed-effects model analysis.

General discussion

Previous research revealed inconsistent results on the accuracy of metamemory for pictures of naturalistic scenes. Specifically, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) found that MJs are unpredictive of scene memorability, whereas other studies have found that JOLs are moderately predictive of individual memory performance for scene pictures (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). One potential explanation for these discrepant results are differences in the cue basis underlying the two types of metamemory judgements. JOLs might rely more on diagnostic cues than MJs because such cues might be more available when making a judgement for one's own memory during a learning task. Alternatively, methodological differences across studies such as the memory task used for measuring stimulus memorability (i.e., classical old/new recognition memory task versus repeat detection task) or the memory criterion value used for accuracy (i.e., recognition memory performance aggregated across participants versus each participant's own memory performance) might be responsible for the discrepant results. The current study differentiated between these possibilities by systematically investigating the relative accuracy and cue basis of MJs and JOLs for pictures of scenes.

Our three experiments revealed that both MJs and JOLs are moderately accurate at predicting differences in the population memorability of scenes. This finding held across three different measures of judgement resolution (within-subjects gamma correlations, within-subjects Pearson correlations, and a mixed-effects model analysis). Our experiments also revealed that MJs and JOLs have a similar cue basis when pictures differed in aesthetics and interestingness (Experiment 1) or represented a broad range of scene memorability (Experiments 2 and 3). Thus, JOLs and MJs had similar accuracy and cue basis when obtained by similar procedures and differed only in that JOLs referred to people's own memory and were made during a learning task, whereas MJs referred to memorability is a generic item attribute and were made during a judgement-only task. This implies that discrepant findings on the accuracy of JOLs and MJs reported in prior work were largely due to methodological differences across studies.

We also found one crucial difference between MJs and JOLs. That is, MJ accuracy improved considerably when MJs were made after rather than before completing the JOL task. In contrast, this was not true for JOLs. Their accuracy was similar in both task order conditions. Experiment 3 was designed to disentangle which component of the JOL task provides the experiences participants subsequently rely on to make more accurate MJs. Results showed that a learning phase is sufficient for improving MJ accuracy as indicated by all measures of judgement resolution. In addition, Pearson correlations (but not Gamma correlations or a mixed-effects model analysis) indicated that the recognition memory test improved MJ accuracy more than making JOLs. This result is consistent with the finding that MJs were more closely related to normed values of scene memorability after having taken a memory test than after having made JOLs.

Accuracy of MJs and JOLs

Our finding that MJs are predictive of differences in the memorability of scenes at the item level contrasts with Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) results. MJ accuracy was instead consistent with the moderate accuracy of JOLs in our study and other metamemory studies (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). Our MJ results were also in line with Saito et al. (2023), who found that judgements of perceived memorability were predictive of real-world objects and faces memorability. So, evidence is accumulating that people can predict the general memorability of different types of images. This makes it even more interesting to ask for the reasons for the discrepancy in MJ results between our study and Isola et al.'s study.

One potential explanation is that we used a fine-grained judgement scale, while Isola et al. used a binary scale (yes, no). Our participants could therefore make more nuanced scene memorability predictions. However, future research will be needed to test whether the opportunity to make fine-grained distinctions between the memorability of scenes really contributes to MJ resolution. So far, one relevant prior study found that the range of confidence scales does not affect confidence accuracy in a recognition memory task (Tekin & Roediger, 2017).

Another potential explanation for why MJs were accurate in our study but not in Isola et al.'s studies might be that we measured scene memorability in an old/new recognition memory test that followed upon a learning phase, while Isola et al. used a repeat detection task. However, this explanation is inconsistent with two aspects of our results. First, Experiments 2 and 3 showed that MJs and JOLs increased with the normed values of scene memorability obtained in Isola et al.'s detection task. Second, relating MJs with normed values of scene memorability revealed very similar results as did relating MJs with the population scene memorability measure obtained in this study. These observations suggest that MJ accuracy is similar for classical old/new recognition memory tasks and repeat detection task.

Regarding the criterion for accuracy, MJs had similar moderate accuracy as JOLs at the item and individual level in our study and other metamemory studies (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This suggests that the lack of MJ accuracy reported in Isola et al. was not due to analysing accuracy at the item level. This is not to say, however, that differences in accuracy between the item and individual level cannot exist. Quite to the contrary, idiosyncratic influences on memory and metamemory that can only contribute to judgement predictive accuracy at the individual level have been obtained in several studies (see, for example, Tullis & Fraundorf, 2017; Undorf et al., 2022).

Order effects on MJ accuracy

Experiment 3 finding that a learning phase improves MJ accuracy suggests that having seen and intentionally learned pictures for oneself provides a good basis for assessing the general memorability of pictures. Interestingly, we did not find evidence that making JOLs *per* se improved MJ accuracy. This is in line with West et al. (2023), who showed that the well-documented increase in JOL accuracy through repeated trials does not rely on making JOLs. Thus, evidence so far indicates that experience with making metamemory judgements *per* se is not essential for subsequent metamemory accuracy.

Regarding the individual contribution of the memory test on MJ accuracy, Pearson correlations showed that a recognition memory test enhances MJ accuracy relative to merely having made JOLs. This finding is consistent with positive effects of test experience on metamemory accuracy reported in studies using the same verbal materials across multiple study-test cycles (Finn & Metcalfe, 2008; Hertzog et al., 2013; King et al., 1980; Koriat & Bjork, 2006a; Touron et al., 2010; but see Mitton & Fiacconi, 2020). However, it should be considered with caution, because it did not replicate in analyses based on gamma correlations or linear mixed models. Nevertheless, in the current study with different material across trials, prior testing experience reliably increased the cue sensitivity of metamemory judgements for a new set of structurally similar scenes. This implies that participants extracted information diagnostic of memorability from testing and used this information to subsequently judge the general memorability of scenes.

To sum up, participants learned about scene memorability by experience with their own learning and testing. This illustrates what Flavell (1979) suggested in his seminal work about metacognition by saying that experiences can "affect the metacognitive knowledge base by adding to it, deleting from it, or revising it" (p. 908).

Future research directions

Given the finding that both JOLs and MJs are predictive of scene memorability, it is important to ask if participants are aware of image features diagnostic of memorability. Based on metamemory research, it is likely that some cue information reaches the level of conscious awareness (e.g., Mueller et al., 2013, 2014). However, it is also plausible that some cues remain experiential at the level of subjective feelings that may not be fully articulated, but nevertheless serve as an inferential basis for the metamemory judgements (e.g., Besken, 2016; Koriat & Levy-Sadot, 1999; Undorf et al., 2017). Prominent inferential accounts of metamemory (Koriat, 1997), distinguish between two types of processes through which cues affect metamemory judgements: theory-based and experience-based processes. Theory-based processes imply the deliberate application of explicit beliefs and knowledge about memory in general and one's own memory. In contrast, experience-based processes imply a non-analytic inferential process that operates below full awareness through which by-products of the cognitive processing of items such as the feeling of "ease" influence metamemory judgements.

Shedding light on participants awareness of stimulus memorability by examining the contributions of theory-based and experience-based processes on metamemory judgements for scene pictures would be an interesting avenue for future research. For instance, this could be done by soliciting pre-study metamemory judgements or using survey designs for assessing the contributions of beliefs to metamemory judgements about item memorability in general.

Limitations

A limitation of Experiment 3 is that we did not include a control group that completed the MJ task twice. We thus cannot fully exclude the possibility that experience with materials during an MJ task might be sufficient to increase MJ accuracy on a second trial. We do, however, regard it unlikely because completing the MJ task did not improve JOL accuracy. Nevertheless, more research will be needed to test whether completing the MJ task repeatedly improves accuracy and if so, whether the improvement is comparable to the one observed after learning pictures for oneself.

Another limitation is that our MJ task was not fully identical to Isola et al.'s task. Our aim was to examine whether differences in accuracy between JOLs and MJs were due to differences in their cue basis arising from the different aspects of memorability judged (one's own vs. generic item attribute) in different tasks (during learning vs. judgement-only). For this, it was necessary to make their procedures similar in all other respects. A potential drawback of this approach is that we cannot know which procedural change or combination of procedural changes are responsible for the differences in MJ accuracy obtained in Isola et al.'s study and the current study.

Conclusion

In conclusion, we found that the predictive accuracy of MJs is not necessarily different from that of JOLs. This stands in stark contrast to Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) findings but is consistent with evidence that metamemory for scene pictures is moderately accurate (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). Our work shows that people can predict not only their own future memory performance for scene pictures but also the general memorability of scene pictures with moderate accuracy. At the same time, we did find a notable difference between JOLs and MJs: MJ accuracy improves with prior learning and testing experience, whereas JOL accuracy is independent of prior assessments of general memorability. This shows that reflections about and experiences with one's own learning and memory contribute to our understanding and knowledge about metamemory and memory processes in general.

Acknowledgements

The authors thank Franziska Schäfer and Anna Fee Wefelmeier for help with material preparation.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Grants UN 345/2-1 (M.U.) and BR 2130/14-1 (A.B.) from the Deutsche

Forschungsgemeinschaft, by the University of Mannheim's Graduate School of Economic and Social Sciences (S.N.-B.) and by a Margarete von Wrangell fellowship from the state of Baden-Württemberg (M.U.).

ORCID iDs

Sofia Navarro-Báez https://orcid.org/0000-0002-6467-654X
Monika Undorf https://orcid.org/0000-0002-0118-824X

Data accessibility statement

The data and materials from the present experiment are publicly available at the Open Science Framework website: https://osf.io/hpy6q/. Experiments 1 and 2 were not preregistered. Experiment 3 was preregistered at https://osf.io/3fujm.

Supplementary material

The supplementary material is available at qjep.sagepub.com.

Notes

1. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) used uncorrected hit rates as a measure of scene memorability. We think that this is not legitimate in our experiments because FA rates ranged between 9% (Experiment 1) and 16% (Experiment 3). Please note that the use of corrected hit rates is a deviation from the preregistration of Experiment 3. Importantly, all results were identical when using uncorrected hit rates except for the main effect of task in the Pearson correlation analysis in Experiment 1 and the interactive effect in the mixed-effects model analysis in Experiment 3.
2. It was impossible to investigate the relative accuracy of MJs with respect to the participant's own memory performance because participants did not complete a recognition memory test on MJ items.
3. Means and (SDs) of aesthetics and interestingness, respectively, were: 0.13 (0.07) versus 0.49 (0.11) for scenes low in aesthetics and interestingness, 0.14 (0.07) versus 0.83 (0.06) for scenes low in aesthetics and high in interestingness, 0.52 (0.11) versus 0.50 (0.10) for scenes high in aesthetics and low in interestingness, and 0.52 (.10) vs. 0.83 (0.05) for scenes high in aesthetics and interestingness.
4. Please note that low memory was a valid reason for rejection on Prolific when we collected data for Experiment 1 in 2020 (Prolific guidelines have changed in this respect in the meantime).
5. Hit rates revealed the same pattern, aesthetics: $F(1, 50) = 22.71$, $p < .001$, $\eta_p^2 = .31$, interestingness: $F(1, 50) = 11.19$, $p < .01$, $\eta_p^2 = .18$, no other effects were significant, $F <= 2.96$, $p >= .09$.
6. The MJ task numerically improved JOL resolution, but the effect was not reliable and only half the size of that for MJs.
7. Means and SDs of each of level of scene memorability were 27.63 and 4.96 (Level 1), 38.72 and 2.72 (Level 2), 44.71 and 1.38 (Level 3), 48.98 and 1.42 (Level 4), 53.80 and 1.42 (Level 5), 58.22 and 1.38 (Level 6), 62.95 and 1.52 (Level 7), 67.75 and 1.52 (Level 8), 72.52 and 1.60 (Level 9), 83.04 and 4.55 (Level 10).

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1417–1433. https://doi.org/10.1037/xlm0000246

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, *197*, 153–165. https://doi.org/10.1016/j.actpsy.2019.04.011

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Caplan, J. B., Sommer, T., Madan, C. R., & Fujiwara, E. (2019). Reduced associative memory for negative information: Impact of confidence and interactive imagery during study. *Cognition and Emotion*, *33*(8), 1745–1753. https://doi.org/10.1080/02699931.2019.1602028

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*(3), 273–280. https://doi.org/10.3758/BF03200854

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. https://doi.org/10.3758/BF03210921

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, *33*(4), 545–565. https://doi.org/10.1006/jmla.1994.1026

Dunlosky, J., & Thiede, K. W. (2013). *Metamemory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019

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

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 1–29. https://doi.org/10.1167/7.1.10

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

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

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology*, *1*(1), 23–32. https://doi.org/10.1037/arc0000003

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, *87*(4), 398–410. https://doi.org/10.1037/0033-295X.87.4.398

Hourihan, K. L. (2020). Misleading emotions: Judgments of learning overestimate recognition of negative and positive emotional images. *Cognition and Emotion*, *34*(4), 771–782. https://doi.org/10.1080/02699931.2019.1682972

Hourihan, K. L., & Bursey, E. (2017). A misleading feeling of happiness: Metamemory for positive emotional and neutral pictures. *Memory*, *25*(1), 35–43. https://doi.org/10.1080/09658211.2015.1122809

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. *Advances in Neural Information Processing Systems*, *24*, 2429–2437. https://doi.org/10.1167/12.9.1082

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

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In 24th IEEE conference on computer vision and pattern recognition (CVPR) (pp. 145–152).

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. https://doi.org/10.1037/a0025960

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. E. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience*, *8*(12), 1776–1783. https://doi.org/10.1038/nn1595

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92–107. https://doi.org/10.3758/BF03211579

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, *93*(2), 329–343. https://doi.org/10.2307/1422236

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments. *Process Theories*, *20*, 483–502.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 595–608. https://doi.org/10.1037/0278-7393.32.3.595

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464–470. https://doi.org/10.1037/0278-7393.16.3.464

Lin, Q., Yousif, S. R., Chun, M. M., & Scholl, B. J. (2021). Visual memorability in the absence of semantic content. *Cognition*, *212*, 104714. https://doi.org/10.1016/j.cognition.2021.104714

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

Mitton, E. E., & Fiacconi, C. M. (2020). Learning from (test) experience: Testing without feedback promotes metacognitive sensitivity to near-perfect recognition memory. *Zeitschrift für Psychologie*, *228*(4), 264–277. https://doi.org/10.1027/2151-2604/a000424

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (*Vol. 26*, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(2), 155–160. https://doi.org/10.1037/h0082899

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 20. https://doi.org/10.5334/irsp.181

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency—Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012

Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language*, *94*, 177–194. https://doi.org/10.1016/j.jml.2016.12.003

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Saito, J. M., Kolisnyk, M., & Fukuda, K. (2023). Judgments of learning reveal conscious access to stimulus memorability. *Psychonomic Bulletin & Review*, *30*, 317–330. https://doi.org/10.3758/s13423-022-02166-1

Schmoeger, M., Deckert, M., Loos, E., & Willinger, U. (2020). How influenceable is our metamemory for pictorial material? The impact of framing and emotionality on metamemory judgments. *Cognition*, *195*(104112), 1–10. https://doi.org/10.1016/j.cognition.2019.104112

Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2004). Computational modeling of face recognition based on psychophysical experiments. *Swiss Journal of Psychology*, *63*(3), 207–215. https://doi.org/10.1024/1421-0185.63.3.207

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2014). Measuring memory and metamemory. In *Handbook of metamemory and memory*. Routledge. https://doi.org/10.4324/9780203805503.ch6

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Tauber, S. K., Dunlosky, J., Urry, H. L., & Opitz, P. C. (2017). The effects of emotion on younger and older adults' monitoring of learning. *Aging, Neuropsychology, and Cognition*, *24*(5), 555–574. https://doi.org/10.1080/13825585.2016.1227423

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 49. https://doi.org/10.1186/s41235-017-0086-z

Touron, D. R., Hertzog, C., & Speagle, J. Z. (2010). Subjective learning discounts test type: Evidence from an associative learning and transfer task. *Experimental Psychology*, *57*(5), 327–337. https://doi.org/10.1027/1618-3169/a000039

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Undorf, M. (2020). Fluency illusions in metamemory. In A. M. Cleary & B. L. Schwartz (Eds.), *Memory quirks* (1st ed., pp. 150–174). Routledge. https://doi.org/10.4324/9780429264498-12

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

Undorf, M., & Bröder, A. (2021). Metamemory for pictures of naturalistic scenes: Assessment of accuracy and cue utilization. *Memory & Cognition*, *49*(7), 1405–1422. https://doi.org/10.3758/s13421-021-01170-5

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me"—Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*(105011), 1–9. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Navarro-Báez, S., & Zimdahl, M. F. (2022). Metacognitive illusions. In R. F. Pohl (Ed.), *Cognitive illusions* (3rd ed., pp. 307–323). Routledge. https://doi.org/10.4324/9781003154730-22

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, *46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6

Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, *92*, 293–304. https://doi.org/10.1016/j.jml.2016.07.003

West, J. T., Kuhns, J. M., Touron, D. R., & Mulligan, N. W. (under review). *Increased metamemory accuracy with practice does not require practice with metamemory*.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). https://doi.org/10.1109/CVPR.2010.5539970