Cognition 240 (2023) 105539

Contents lists available at ScienceDirect

Cognition

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

Conscious and unconscious memory and eye movements in context-guided visual search: A computational and experimental reassessment of Ramey, Yonelinas, and Henderson (2019)

Daryl Y.H. Lee [*], David R. Shanks

Department of Experimental Psychology, University College London, UK

ARTICLE INFO

Keywords:
Recognition
Contextual cueing
Computational model
Memory
Visual search
Unconscious learning

ABSTRACT

Are eye movements unconsciously guided towards target locations in familiar scenes? In a recent eyetracking study, Ramey, Yonelinas, and Henderson (2019) measured eye-movement efficiency (scanpath ratio) and memory judgments when participants searched for targets in repeated and novel scenes. When trials judged new with high confidence were selected, scanpath ratio was lower for old scenes (misses) than for new scenes (correct rejections). In addition, familiarity as measured by recognition confidence did not significantly predict scanpath ratio. Ramey et al. attributed these results to unconscious learning guiding eye movements. In a re-assessment of Ramey et al.'s data, we show that their findings can be accounted for by a single-system computational model in which eye movements and memory judgments are driven by a common latent memory representation. In particular, (a) the scanpath ratio difference between high-confidence misses and correct rejections is a consequence of regression to the mean, while (b) the null correlation between familiarity and scanpath ratio, partly a natural consequence of the low reliability of the scanpath ratio measure, is also reproduced by the model. Two pre-registered experiments confirm a novel prediction of the alternative single-system model. This work offers a parsimonious account of Ramey et al.'s findings without recourse to unconscious guidance of eye movements.

1. Introduction

A topic of great interest in cognitive psychology concerns the organization of memory. A widespread perspective holds that there are functionally distinct explicit (conscious) and implicit (unconscious) memory systems (Squire, 1992; Squire & Dede, 2015; Tulving & Schacter, 1990). Explicit memory is thought to underpin conscious recall of past experiences, whereas implicit memory is linked to behavioral changes arising from past experiences and is assumed to be inaccessible to awareness (Schacter, 1987). Evidence supporting this *multiple-systems* account has typically demonstrated that performance in explicit and implicit memory tasks can be dissociated. One illustrative dissociation entails a comparison of two tasks measuring priming and recognition respectively (e.g., Jacoby & Dallas, 1981; Richardson-Klavehn, Clarke, & Gardiner, 1999). Priming refers to enhanced performance in behavioral response to a stimulus due to prior exposure; in contrast, recognition refers to the mental faculty of judging whether a stimulus has been presented in a preceding context (i.e., a learning phase; Lange, Berry, & Hollins, 2019).

The multiple-systems perspective, however, has been challenged on methodological and statistical grounds (e.g., Buchner & Wippich, 2000; Dunn, 2003; Poldrack, 1996). For instance, it has been argued that dissociations provide only weak constraints on underlying mechanisms (e.g., Newell & Dunn, 2008), and a number of researchers have presented evidence for an alternative *single-system* view, according to which performance on explicit and implicit tests is driven by a common memory source (e.g., Berry, Kessels, Wester, & Shanks, 2014; Berry, Shanks, & Henson, 2008; Berry, Shanks, Speekenbrink, & Henson, 2012; Lange et al., 2019; Newell, Dunn, & Kalish, 2011; Nosofsky & Zaki, 1998).

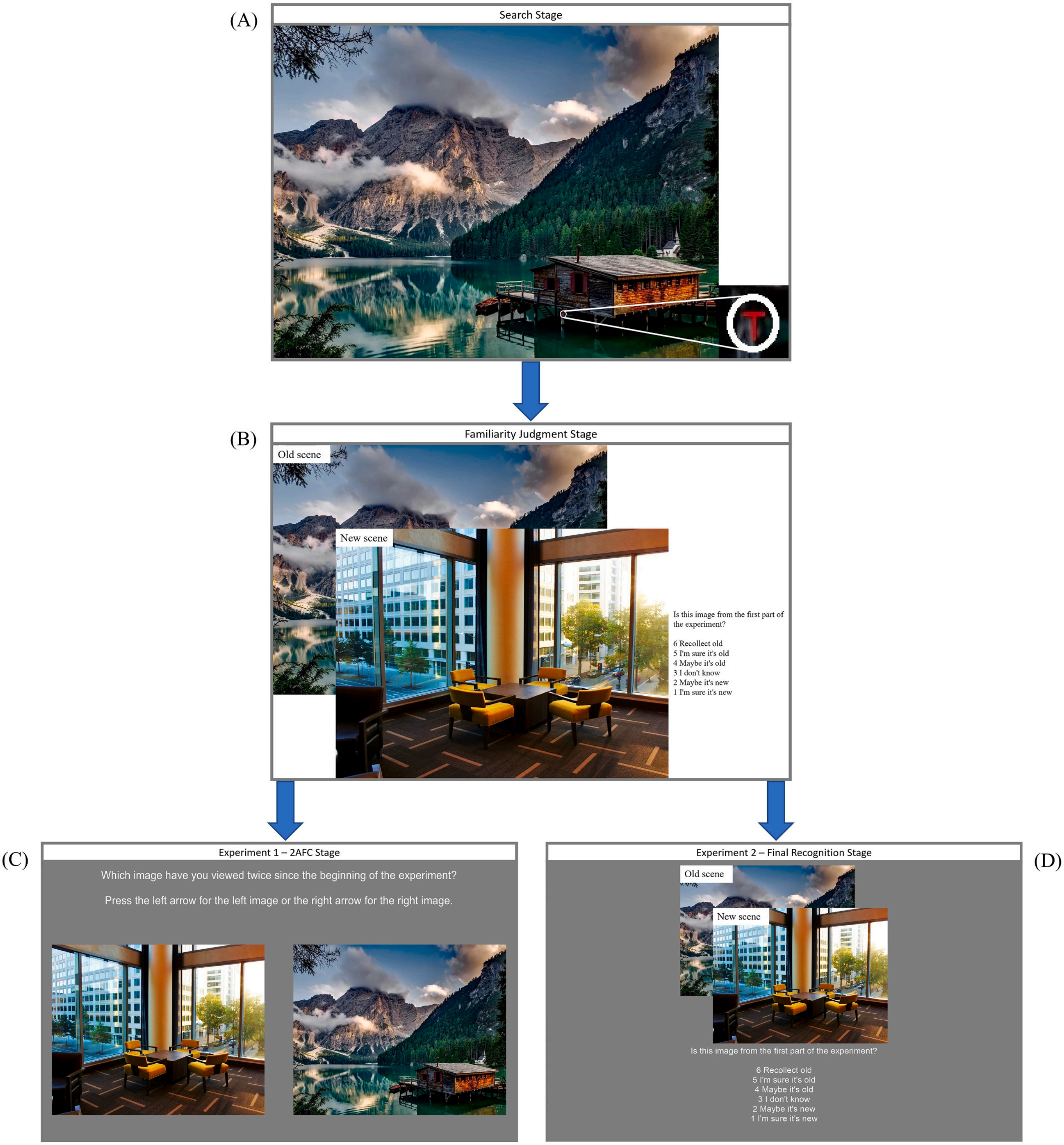
A recent eyetracking study by Ramey, Yonelinas, and Henderson (2019) asked whether eye movements can be unconsciously guided towards target locations in familiar scenes. Their research built on previous attempts to address this issue specifically (Hannula et al., 2010; Hannula, Baym, Warren, & Cohen, 2012; Ryan, Althoff, Whitlow, & Cohen, 2000; Smith & Squire, 2017), as well as work on the more general question of the role of unconscious memory in the guidance of visual attention (see Chun & Jiang, 1998) – all of which have been the

[*] Corresponding author at: UCL Department of Experimental Psychology, 26 Bedford Way, London WC1H 0AP, United Kingdom.
E-mail address: yu.lee.18@ucl.ac.uk (D.Y.H. Lee).

https://doi.org/10.1016/j.cognition.2023.105539
Received 30 June 2022; Received in revised form 20 June 2023; Accepted 26 June 2023
Available online 12 August 2023

subject of considerable controversy (Colagiuri & Livesey, 2016; Jiang, Sisk, & Toh, 2019; Kroell, Schlagbauer, Zinchenko, Müller, & Geyer, 2019; Vadillo, Konstantinidis, & Shanks, 2016; Vadillo, Malejka, Lee, Dienes, & Shanks, 2022).

In Ramey et al.'s (2019) experiment (see Fig. 1 for an overall schematic), participants first saw a set of 64 real-world scenes each presented once alongside another set of 64 scenes that were presented three times in the learning phase. The scenes presented once were included because, as noted by the authors, it was anticipated that the distribution of recognition responses to scenes presented three times could be skewed due to ceiling memory performance. Either a small letter 'L' or 'T', the search target, was superimposed on each scene, and participants were


(A)
Search Stage
(B)
Familiarity Judgment Stage
Old scene
New scene
(C)
Experiment 1 – 2AFC Stage
(D)
Experiment 2 – Final Recognition Stage
Old scene
New scene

Is this image from the first part of the experiment?

6 Recollect old
5 I'm sure it's old
4 Maybe it's old
3 I don't know
2 Maybe it's new
1 I'm sure it's new

Which image have you viewed twice since the beginning of the experiment?

Press the left arrow for the left image or the right arrow for the right image.

Is this image from the first part of the experiment?

6 Recollect old
5 I'm sure it's old
4 Maybe it's old
3 I don't know
2 Maybe it's new
1 I'm sure it's new

Fig. 1. Schematic diagram of Ramey et al.'s (2019) task and Experiments 1 and 2. (A) In the learning phase participants searched for a target and responded to its identity. Here the target is magnified for illustration. (B) In the familiarity judgment stage (i.e., 'test phase'), a preview of an old or new scene without its target was presented for 0.4 s at the beginning of each trial, which was then replaced by a recognition confidence rating screen. After the participant gave a rating, the same scene with its target was presented and participants searched for the target as in the first stage, with eye movements being monitored. (C) The additional 2AFC test stage in Experiment 1. (D) The additional single-item test stage in Experiment 2.

required to find and identify the target (Fig. 1A). In a subsequent test phase, a total of 160 scenes were presented. Of these, 128 were familiar ones previously presented during the learning phase, while the remaining 32 were new scenes serving as lures. In each trial (Fig. 1B), participants previewed the scene without its target and then indicated whether they recalled viewing it during the study phase based on a 6-point recognition confidence scale, ranging from 1 = *I'm sure it's new* to 5 = *I'm sure it's old*, with 6 = *Recollect old* as an additional response designed to capture judgments putatively driven by a conscious recollection process distinct from familiarity-based memory (Yonelinas, 1994, 2002). Then, the same scene (with the search target this time) was presented again, and participants searched for and responded to the target as they did in the previous learning phase. In both phases, eye movements were monitored.

In order to investigate eye movements during target search, special emphasis was put on the scanpath ratio (*SPR*) as a measure of search efficiency. Scanpath ratio is defined as the ratio of the total distance travelled by the eyes when finding the target to the shortest possible distance from the fixation point to the target. Optimal performance is indicated by a scanpath ratio of 1, whereas high ratios indicate that inefficient paths have been taken. Focusing predominantly on data for the new scenes and those presented once in the learning phase, Ramey et al. (2019) found, among other things, that there was no significant difference in scanpath ratio in the test phase between old items that were recollected (i.e., with ratings of 6) and old items that were recognized with the highest level of familiarity-based confidence (i.e., with ratings of 5). Furthermore, familiarity strength (excluding recollection) in terms of recognition confidence (across ratings of 1–5) did not significantly predict the scanpath ratio. Even more strikingly, among scenes judged as 'sure new' – that is, scenes for which no conscious memory was detectable – scanpath ratio was found to be significantly lower for old scenes (i.e., high-confidence misses) than for new scenes (i.e., high-confidence correct rejections), as shown in Fig. 2. The same patterns of results were observed in analyses on old scenes presented three times in the learning phase. Ramey et al. (2019) presented evidence that improved scanpath performance involved general improvement in efficiency across all individual saccades in a search trial (the modal number of saccades was 7), rather than being driven by more efficient early saccades or fewer saccades in each trial. Importantly, Ramey et al. (2019) attributed these findings to facilitation by unconscious memory in scanpath efficiency of target search in repeated scenes.

Are these findings truly evidence of unconscious memory? We address this question by formulating a model which assumes that a single latent memory representation determines both memory judgments and eye movements. If this model is able to capture the key findings obtained by Ramey et al. (2019), then their interpretation is challenged.

2. A single-system model of eye movements and memory judgments

2.1. Model specification

In the following, we present a single-system model and a *post hoc* model-fitting simulation. The single-system model follows a long tradition in assuming that a single latent source may drive both unconscious (implicit) and conscious (explicit) behavioral measures (Berry et al., 2008; Berry et al., 2012; Jamieson, Holmes, & Mewhort, 2010; Schimmack, 2021; Shanks & Berry, 2012; Shanks & Perruchet, 2002; Vadillo et al., 2022; Zaki & Nosofsky, 2001). It is conceptually grounded in signal-detection theory (SDT; Green & Swets, 1988). The model assumes that both the performance measure (i.e., scanpath ratio) and the recognition confidence measure are driven by a common underlying memory representation, such that responding to each scene presented during the test phase is determined by a random, normally distributed latent memory strength variable *S*. In light of previous exposure during the learning phase, old scenes are on average associated with higher *S* values than new scenes.

Ramey et al.'s (2019) data allow for a direct estimation of the magnitude of *S* for old and new scenes. Specifically, all correct recognitions of old scenes presented once at study were treated as hits (i.e., responses receiving a confidence rating between 4 and 6), whereas all incorrect endorsements of new scenes as old ones were treated as false alarms (again, responses receiving a confidence rating between 4 and 6). This yielded a mean d' score of 1.373, which represents participants' mean performance in the recognition task. In order to model the d' score, in the simulation, *S* takes a mean value of 1.4 for old scenes and 0 for new scenes, while the standard deviation (*SD*) of *S* for both types of scenes takes a value of 0.1. Furthermore, to generate recognition responses, a random, normally distributed error term (e_{REC}) with a mean of 0 and a *SD* of 1 is added to *S* to generate a continuous recognition variable (*REC*):

$$REC = S + e_{REC} \tag{1}$$

These imputed parameter values ensure that the required d' value is preserved. For the simulation results reported below, the ensuing value of d' is 1.376. We simulate a total of 1000 participants, each responding to 64 old (presented once at study) and 32 new scenes, identical to the number of trials of the respective scene types in Ramey et al.'s (2019) experiment. Since the underlying memory strength is likely to be different from one scene to another even if both scenes are of the same type, a value of *S* and a value of e_{REC} are sampled afresh from the

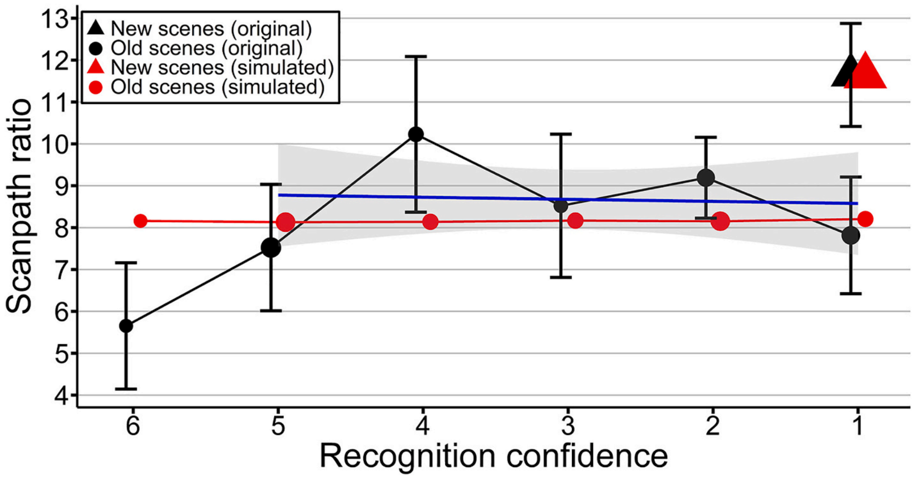


Fig. 2. Mean scanpath ratio (*SPR*) across different levels of recognition confidence during the test phase in Ramey et al.'s (2019) experiment, and simulation results. The blue regression line is the prediction of *SPR* across different levels of recognition confidence (excluding ratings of 6) for old scenes using the original data. Symbol sizes represent the relative proportions of scenes across different levels of recognition confidence. The shaded area represents the 95% confidence band of the regression line, and the error bars represent the 95% confidence intervals of the *SPR* means. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Best-fitting parameter values.

Parameter		Value	Remark
S (old)	M	1.4	d' is equal to mean S (old)/ $SD(REC)$. Because $SD(REC) = \sqrt{SD(S)^2 + SD(e_{REC})^2} = \sqrt{0.1^2 + 1^2} = 1.005$, the value of mean S (old) ensures the appropriate d' value.
	SD	0.1	
S (new)	M	0	
	SD	0.1	
e_{REC}	M	0	Default values.
	SD	1	
e_{SPR}	M	0	
	SD	1	
C_1		−0.48	The parameter values are z-transformed overall cumulative proportions of ratings 1–6, computed by taking the inverse of the cumulative distribution function of the generated REC values. They distribute each participant's responses into 6 recognition bins in the same frequencies as in the behavioral data.
C_2		0.09	
C_3		0.41	
C_4		0.74	
C_5		1.50	
a		38.10	Best-fitting parameter values based on maximum likelihood estimation.
b		3.29	

−2 log-likelihood = 25.24; Akaike Information Criterion = 29.24; Bayesian Information Criterion = 29.13.

corresponding distributions for each trial, and then a value of REC is computed based on Eq. (1).

As reported by Ramey et al. (2019), the proportions of old scenes receiving ratings of 1–6 were, respectively, 20%, 19%, 12%, 14%, 24%, and 11%; in contrast, for new scenes, the proportions were, respectively, 54%, 25%, 13%, 5%, 2.5%, and 0.5%. Collapsed across the two types of scenes, the overall proportions of ratings 1–6 were, respectively, 31.54%, 22.03%, 12.29%, 11.14%, 16.28%, and 6.72%. Accordingly, we computed five REC cut-offs based on the overall cumulative proportions of the six ratings (defined by $C_1 - C_5$; see Table 1) and then split the REC values generated for each simulated participant into 6 bins, corresponding to the 6-point recognition confidence scale.

Because both recognition and SPR are assumed to be driven by a common memory source, for each scene, the same value of S is used to generate its recognition rating and SPR. Importantly, in light of the differences in task requirements between recognizing a scene and finding a search target, it is further assumed that an independent source of error reflecting non-memorial noise (e_{SPR}) contributes to SPR. As such, the value of S for a given scene is combined with a different error term, which is normally distributed with a mean of 0 and a SD of 1. These parameters are the same as those assumed for e_{REC}. Since smaller SPR values indicate better performance, SPR is assumed to be an inverse function of S:

$$SPR = \frac{a}{b + S + e_{SPR}} \tag{2}$$

where parameters a and b serve merely as scaling factors to ensure that SPR values are in the observed range of scanpath ratios. It then becomes possible to test whether the pattern of results obtained by Ramey et al. (2019) can be reproduced based on the simulated REC and SPR data. Code for the simulation is available at https://osf.io/8nfqj/.

The parameters of the model can be grouped roughly into three sets. One set (including mean S (new) and the means and SDs of e_{SPR} and e_{REC}) are simple default values. The second set includes parameters whose values are determined by general aspects of the observed data. This set includes mean S (old), whose value determines the obtained overall recognition d' value, and $C_1 - C_5$, whose values are computed after the raw simulated data are generated in order to distribute responses across the 6 recognition bins in the correct frequencies. Lastly, parameters a and b (Eq. 2) are free parameters whose values are determined by maximum likelihood, and the SD of S (old) = SD of S (new) was set by manual trial-and-error. The best-fitting parameter values are shown in Table 1. The simulation results reported below depend heavily on the SDs of S (old) and S (new) being small relative to the SDs of the error terms. We assess the sensitivity of the results to the chosen parameter values below.

In addition to the scanpath ratio, Ramey et al. (2019) also assessed first saccade accuracy (FSA), the angular (degree) error with which the first eye movement in a trial was directed towards the target. A lower degree error value indicates that the first saccade was more accurate. Unlike the scanpath ratio results, Ramey et al. (2019) did not find a significant difference in FSA between old and new scenes given the lowest recognition confidence rating, though, like scanpath ratio, FSA was also not significantly predicted by familiarity strength. As Ramey et al. (2019) did not attribute the FSA findings to unconscious memory, we do not attempt to model this dependent variable, save that we comment on the contrast between the FSA and SPR results below.

2.2. Simulation results

The model was fitted to the data using maximum likelihood parameter estimation, based on Nelder and Mead's (1965) Simplex algorithm as implemented in R (R Core Team, 2022). The fitted data comprised all the points in Fig. 2 – that is, the SPR values for old scenes given recognition ratings of 1–6 and new scenes given a rating of 1; SPR values for new scenes given ratings of 2–6 were excluded, as in Ramey et al.'s (2019) analyses. Specifically, there were seven data points to be fitted averaged across a total of 1400 trials from 23 participants in Ramey et al.'s experiment, after excluding invalid trials. On a number of trials response times (RTs) over 20 s were recorded due to eyetracking software error, even though all trials should have been terminated at 20 s without response. In these trials, participants did not give any response to identify the target letters. We decided to treat such trials as invalid, even if scanpath ratio data were recorded, as it was not clear whether the software error would affect the scanpath ratio data or whether participants performed the search for target letters in these trials. Additionally, trials with missing scanpath ratio data were also excluded.

The simulation results are shown in Fig. 2, and the model parameters are reported in Table 1.

Note that the observed SPR data in Fig. 2 are slightly different from those reported by Ramey et al. (2019), which were estimated marginal means derived from their linear mixed effects models.

It is striking that the simulated SPR results across all recognition confidence levels approximate the original findings quite closely. Crucially, a difference in mean SPR values between old and new scenes at the lowest confidence level (i.e., judgment of 'sure new') is evident, even though both recognition and SPR are based on the same latent memory variable S. The simulated data also reveal a negligible association between familiarity (ratings 1–5) and SPR. In other words, enhancement in scanpath efficiency for unrecognized scenes is present, as well as minimal association between familiarity and SPR, even though the model does not distinguish between conscious and unconscious memory.

To evaluate the sensitivity of the results to the specific parameter

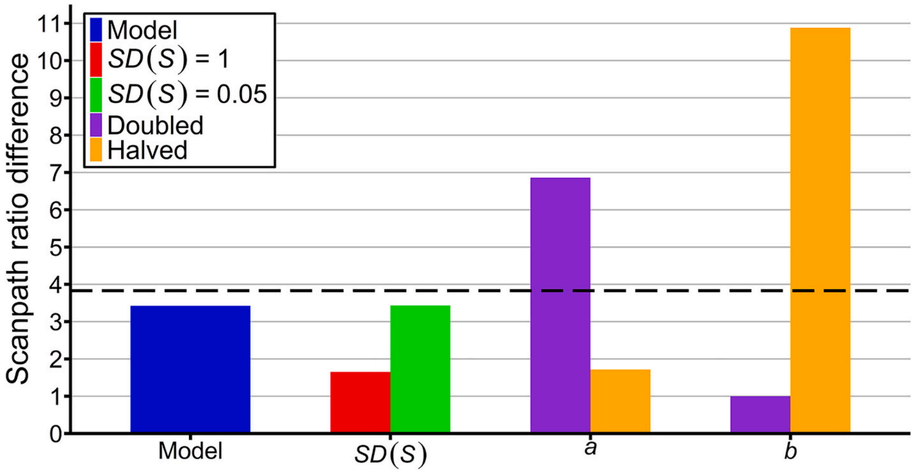


Fig. 3. Mean simulated scanpath ratio (*SPR*) difference between old and new scenes rated 1 across variations in parameter values. The blue bar represents the difference based on the best-fitting parameter values. Each of the remaining bars represents the difference when the specified parameter change is made, holding all others at their values in Table 1. The dashed line represents the *SPR* difference in the behavioral data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

values, we focus on the key comparison of new versus old scenes endorsed as ‘sure new’ (i.e., ratings of 1) in terms of mean *SPR*. As illustrated in Fig. 3, the model is remarkably robust in producing the *SPR* difference in the expected direction (i.e., higher *SPR* for new than old scenes) across a range of parameter values.

Why does the model generate these patterns? We explore this question in the following sections.

3. Regression to the mean and *post hoc* scene selection

It is reasonable to assume, and indeed is a fundamental tenet of classical test theory (Hambleton & Jones, 1993), that when an aggregated set of data is analyzed, errors cancel out resulting in mean error of zero. However, the same cannot be said for a selected subset of data. It is easy to see why this is the case by re-examining Eq. (1). Here, although S and e_{REC} are independent and thus uncorrelated, observed *REC* values are necessarily correlated with e_{REC} as they share a common term (i.e., e_{REC} itself). In light of this positive correlation, higher observed *REC* values on average incorporate higher e_{REC} values, and vice versa.

It follows that observed *REC* values of scenes selected *post hoc* are susceptible to systematic bias induced by non-zero error components on average (Rothkirch, Shanks, & Hesselmann, 2022; Shanks, 2017). Whereas the old/new status of a test item is determined in advance by the experimenter, assignment to the high-confidence miss or correct rejection categories occurs *post hoc*, depending on the participant’s decision. This fact is fundamental to the model’s behavior. Recall that the distribution of S values for old scenes is much higher than for new scenes. On average, any given *REC* value of a ‘sure new’ old scene is likely to comprise an underlying memory representation value S that is not particularly extreme combined with an extreme error term (i.e., a large negative e_{REC} value). But because e_{REC} and e_{SPR} are independent, it is improbable that the same underlying S value will be combined with a e_{SPR} value that is as extreme as e_{REC}; hence regression to the *SPR* mean for old scenes will occur, leading to a relatively low *SPR* value. On the other hand, any *REC* value for a new scene judged ‘sure new’ will tend to be made up of a less extreme e_{REC} value. When its underlying S value is in turn combined with an independent e_{SPR} value, regression to the mean will again occur (this time to the mean *SPR* for new items), yielding a relatively high *SPR* value. Indeed, this is the pattern that emerges in the current simulation, as shown in Fig. 4 which decomposes the simulated recognition ratings into their constituent parts S and e_{REC}.

Fig. 4 shows several noteworthy patterns, all inevitable statistical consequences of the model equations and *post hoc* selection. First, while mean error is zero overall for both scene types, the errors are not zero within each recognition confidence category: this is the essence of the bias introduced by *post hoc* selection (see also Rothkirch et al., 2022; Shanks, 2017). Secondly, while it is intrinsic to the model that mean S is greater for old than new scenes, this is also the case within any given bin (red lines). Just because old and new scenes might evoke the same recognition judgment, this does not entail that the latent memory strengths underlying them are identical. Instead, they evoke similar judgments because *post hoc* selection ensures that within each bin, e_{REC} is larger for new than old scenes (blue lines). Thirdly, if we focus on the crucial recognition confidence = 1 category, it can be seen that the description given in the previous paragraph is borne out: old scenes in this category have extreme negative values of e_{REC}. The values of S for old and new scenes in this category are divergent (just as for all other categories), and this difference explains the *SPR* difference for old and new scenes.

It follows that a difference in *SPR* values between the two types of scenes selected *post hoc* is bound to emerge due to regression to the mean, which occurs for any bivariate data based on imperfectly correlated measures (Campbell & Kenny, 1999; Mee & Chua, 1991). As such, the *post hoc* data selection method employed by Ramey et al. (2019) is ill-suited for demonstrating unconscious memory, unless SPR and recognition confidence are perfectly correlated and not subject to measurement error. Needless to say, these criteria are virtually impossible to meet.

4. A novel test of the model: Reversing the variables

To further illustrate that Ramey et al.’s (2019) finding is likely a statistical inevitability due to *post hoc* scene selection from noisy bivariate data (Campbell & Kenny, 1999; Rothkirch et al., 2022; Shanks, 2017), we ran an additional analysis on their data by turning the dependent and independent variables around. The logic of this analysis is that as far as the model is concerned, recognition confidence and SPR are two parallel measures of a common latent memory representation, and so regression to the mean is expected to occur not only when SPR is conditionalized on recognition confidence (Fig. 2) but also when recognition confidence is conditionalized on SPR.

Therefore, we divided each participant’s *SPR* values, collapsed across old and new scenes, into 6 equal *SPR* bins (i.e., sextiles). We then calculated the mean recognition confidence ratings for old and new scenes across all *SPR* bins. These data are shown in Fig. 5. Crucially, focusing on the lowest *SPR* bin (i.e., representing the least learning), the mean recognition confidence rating for old scenes ($M = 3.35$, $SD = 0.71$) is revealed to be significantly higher than that for new scenes ($M = 1.51$, $SD = 0.68$), $t(22) = 13.09$, $p < .001$, $d_z = 2.79$.[1] Fig. 5 also shows that the

[1] The effect size is calculated as $d_z = \frac{t}{\sqrt{df}}$ (Lakens, 2013; Rosenthal, 1991).

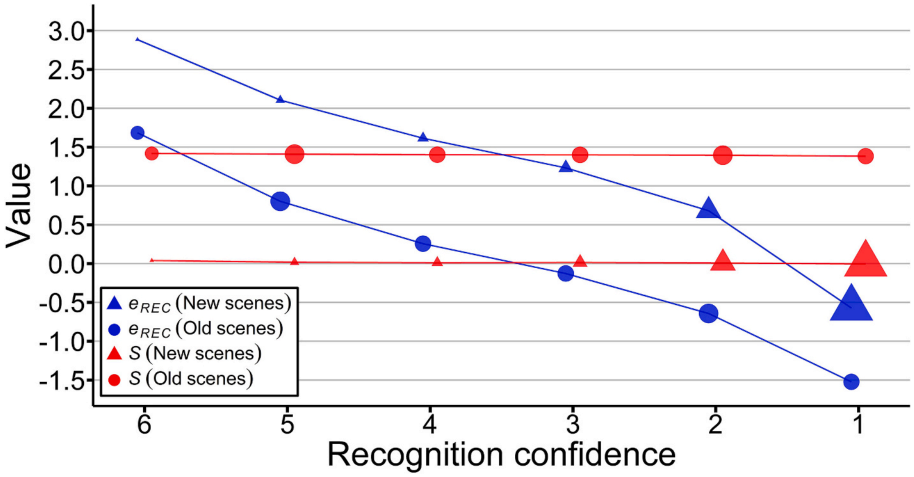


Fig. 4. Mean values of e_{REC} and S for old and new scenes across different levels of recognition confidence in the model simulation. Symbol sizes represent relative proportions of scenes across different levels of recognition confidence. The proportions of old scenes receiving ratings of 1–6 were, respectively, 14.86%, 21.92%, 15.32%, 15.14%, 22.09%, and 10.68%; in contrast, for new scenes, the proportions were, respectively, 64.04%, 21.79%, 6.85%, 4.10%, 2.70%, and 0.52%.

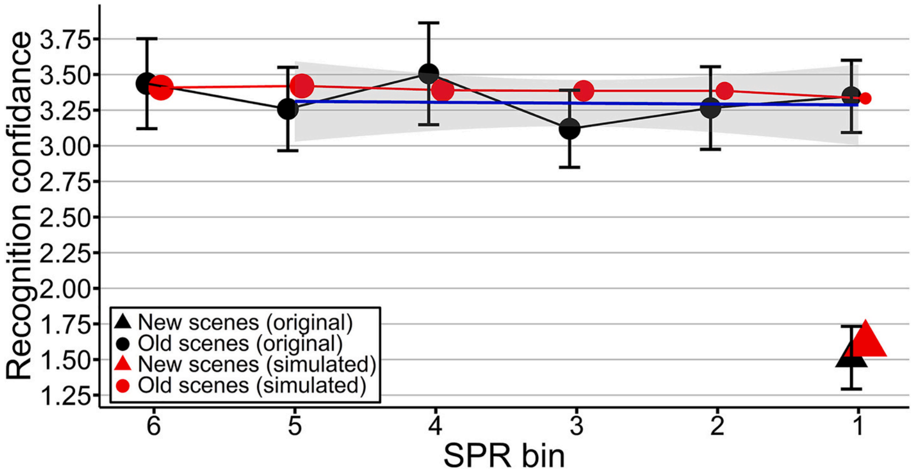


Fig. 5. Mean recognition confidence across different *SPR* bins based on the test phase data in Ramey et al.'s (2019) experiment, and corresponding data from the simulation reported above. The *SPR* bins are arranged in ascending order of search performance (i.e., the *SPR* bin of 6 indicates the most learning, whereas the *SPR* bin of 1 indicates the least learning). The blue regression line is the prediction of recognition confidence across *SPR* bins (excluding bin 6) for old scenes using the original data. The shaded area represents the 95% confidence band of the regression line, and the error bars represent the 95% confidence intervals of the recognition confidence rating means. Symbol sizes represent the relative proportions of scenes across different SPR bins. For the original data, the proportions of old scenes across *SPR* bins 1–6 were, respectively, 14.58%, 16.58%, 17.39%, 15.85%, 17.12%, and 18.48%; for new scenes, the proportions were, respectively, 22.81%, 18.07%, 15.51%, 17.52%, 14.60%, and 11.50%. For the simulated data, the proportions of old scenes across *SPR* bins 1–6 were, respectively, 5.05%, 11.73%, 16.71%, 19.99%, 22.44%, and 24.08%; for new scenes, the proportions were, respectively, 39.90%, 26.53%, 16.58%, 10.03%, 5.12%, and 1.84%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

same pattern emerges when the simulated data are analyzed in the same way: the model generates a difference in mean recognition confidence ratings between old and new scenes at the lowest *SPR* bin, as well as a weak association between recognition confidence and *SPR* bins (excluding the highest *SPR* bin, for ease of comparison with Fig. 2). Indeed, recognition confidence ratings across all *SPR* bins closely approximate those based on Ramey et al.'s (2019) data after running the equivalent analysis. This is not a mere coincidence: insofar as bivariate data are concerned, the model makes the same prediction about regression to the mean regardless of which variable is selected *post hoc*. It does so here for a reason conceptually identical to that described previously and illustrated in Fig. 4. Fig. A.1 in Appendix A.1 shows a breakdown of the S and e_{SPR} values that underlie the simulated data, analogous to Fig. 4. It is striking that this close fit is achieved with the same parameter values chosen to fit the data in Fig. 2.[2]

If we follow Ramey et al.'s (2019) line of reasoning, this result would lead us to the conclusion that recognition judgments are driven by 2 dissociable processes. It goes without saying that such a conclusion is

[2] Although the model was formally fitted to the 7 datapoints in Fig. 2, the additional 7 datapoints in Fig. 5 should be included in any calculation of the total number of datapoints to free parameters. In total the model simulates 14 datapoints via 3 freely-varying parameters, a, b, and SD S (old) = SD S (new).

unparsimonious. Instead of inferring that Ramey et al.'s data provide evidence for a total of 4 distinct processes (2 guiding eye-movements and 2 determining recognition judgments), the model captures all key aspects of the data with a single latent memory variable.

5. Reliability and nonsignificant association

Since the extent of regression to the mean depends on the reliability of the measures concerned (Campbell & Kenny, 1999; Rothkirch et al., 2022; Shanks, 2017), it stands to reason that the less reliable the SPR and recognition confidence measures, the more likely we are to observe lower SPR values for 'sure new' old compared to new scenes; in the limit, if the measurement of recognition is completely unreliable, then the mean SPR values for these scene types at each recognition level will simply equal the grand mean values. Although Ramey et al. (2019) did not provide any reliability estimates regarding their experimental tasks, by analyzing the original test phase dataset, it is possible to estimate the split-half reliabilities of the two measures.

Specifically, all test trials for each participant were randomly split into two equal halves and means were computed for each of the halves. For each measure, the means for the two halves were correlated across participants to serve as a measure of reliability. To minimize the influence of sampling error, we repeated this procedure over 5000 random splits of each participant's data (Parsons, 2020). For the recognition confidence measure, we obtained an uncorrected mean reliability estimate of $r = 0.77$, 95% CI [0.64, 0.88]. After applying the Spearman-Brown correction for attenuation (Brown, 1910; Spearman, 1910), this yielded $r = 0.87$, 95% CI [0.78, 0.94]; in contrast, for the SPR measure, the uncorrected mean reliability estimate obtained was only $r = 0.43$, 95% CI [0.18, 0.68], and the Spearman-Brown corrected estimate was $r = 0.59$, 95% CI [0.30, 0.81].[3] Although the recognition measure has adequate psychometric properties, the reliability of the SPR measure is disappointingly low and falls short of the commonly accepted lower threshold of 0.70 (Savage, 2018). As a consequence, regression to the mean induced by measurement error, particularly in light of the low reliability of the SPR measure, may have been substantial.

Not only is reliability of the measures pertinent to the magnitude of regression to the mean, but it also has implications for interpreting the other major finding reported by Ramey et al. (2019), namely that scanpath ratio was not significantly predicted by familiarity strength for old scenes at test. This is because one measure cannot correlate with another if it does not correlate with itself, which is what reliability refers to. Ramey et al. (2019) reported a non-significant association between scanpath ratio and familiarity strength based on the conventional frequentist approach, as well as moderate Bayesian evidence (i.e., $BF_{10} = 0.14$) for a null relationship between the two variables (Jeffreys, 1961).

Consistent with their frequentist finding, by analyzing the original dataset, we obtained a statistically nonsignificant estimate of the repeated measures correlation between SPR and familiarity, $r = 0.024$, $p = .83$, 95% CI [−0.186, 0.231], after adjusting for between-subjects variance (Bakdash & Marusich, 2017). But as noted above, this null association can be simulated by the single-system model (as in the simulation results in Fig. 2). Therefore, the finding of nonsignificant association between familiarity and scanpath ratio does not provide evidence of true independence at the latent level but may instead be partly ascribed to the low reliability of the SPR measure. In a similar vein, one should also be cautious when interpreting the Bayesian evidence in light of the low reliability of the measure. As demonstrated by Malejka, Vadillo, Dienes, and Shanks (2021), even moderate evidence in favor of the null hypothesis as indicated by a Bayes factor is undermined when uncertainty (e.g., measurement uncertainty) is taken into account, especially if the prior assumes a weak correlation (as opposed to a prior which regards all correlations as equally plausible). This concern might be circumvented in future studies if they include a greater number of test trials to measure SPR with increased reliability.

Whatever its explanation, the key point is that the single-system model also predicts a low recognition/*SPR* correlation as shown in Fig. 2. The mean simulated value, $r = -0.047$, 95% CI [−0.078, −0.016], is within the confidence interval of the observed correlation. Hence this null association is not diagnostic of two latent memory sources.

Note that the reliabilities of the recognition confidence and *SPR* measures are much lower in the simulated data than in the empirical data. The reason for this is straightforward. Reliability depends crucially on between-participants variance (Hedge, Powell, & Sumner, 2018): when this variance is low, reliability is also low. The model does not incorporate such variance – it models each participant as identical, apart from trial-by-trial sampling variation, and with the same mean value of S (old).[4] Under these conditions reliability is not conceptually meaningful.

Systematic between-participants variance can be incorporated into the model by sampling mean S (old) for each participant from a distribution. In Appendix A.2 we describe such a model, which achieves a virtually identical fit to the primary data while also generating recognition confidence and *SPR* measures with reliabilities similar to those in the behavioral data. However this is achieved at the cost of making the model more complex and adding to the number of parameters.

6. Do different eye movement measures dissociate conscious and unconscious memory?

We noted earlier that Ramey et al. (2019) analyzed a second eye-movement measure as a function of memory responses, namely FSA. We did not attempt to model this dependent measure because our focus is on unconscious drivers of eye movements, and this dependent variable did not show a difference between old and new scenes judged 'sure new'. However, another finding from this measure deserves comment. Ramey et al. (2019) found that FSA (but not scanpath ratio) as measured by degree error of the first saccade was significantly smaller for scenes participants recollected (i.e., a rating of 6) compared to those rated as 'sure old' (i.e., a rating of 5). By decomposing eye movements on these trials, Ramey et al. (2019) also presented some evidence that this conscious recollection effect caused a subset of first saccades to be highly accurate, rather than causing a more incremental improvement in FSA across all recollected trials. Ramey et al. (2019) noted that a feature of the task may have contributed to this pattern – namely that in the test stage, each scene was previewed prior to target search, meaning that conscious recollection could have affected planning for target search prior to the eye movement being measured. Be that as it may, in our model, recognition confidence = 6 and recognition confidence = 5 items are not equivalent – the former have higher values of S, and so, depending on the exact process by which FSA is assumed to be related to S, the effect could be a quantitative rather than qualitative one (see Mickes, Wais, & Wixted, 2009, for a discussion of whether recollection is

[3] The reliability estimates were computed based on Ramey et al.'s data after excluding trials in which RTs over 20 s were recorded, as mentioned in Section 2.2 above. If the reliability assessment was conducted using a single odd-even split instead of many random splits, the estimates for the recognition confidence measure would be 0.66 (uncorrected) and 0.80 (corrected) respectively, whereas the estimates for the SPR measure would be 0.42 (uncorrected) and 0.59 (corrected) respectively.

[4] In particular, there is zero between-participants variance in the simulated recognition confidence measure: the mean rating is the same for every participant, which results from the binning of *REC* values based on the same C_1 - C_5 values for each participant. In such an extreme case, any split-half reliability estimate is necessarily −1, as it is equivalent to correlating the difference between the grand mean and the mean of one half of trials with the difference between the grand mean and the mean of the other half of trials. The reliability of the simulated SPR measure is $r = 0.03$, Spearman-Brown corrected.

truly a distinct process or simply represents high-confidence memory). But even if the effect represents a genuine and qualitatively distinct mechanism contributing to eye movements, the key point is that it is a conscious, not unconscious, mechanism.

A more fundamental point raised by Ramey et al. (2019) concerns the relationship between FSA and SPR. They noted that the primary objective of their study was to investigate whether different kinds of eye movement behaviors are influenced selectively by either conscious or unconscious memory. Ramey et al. (2019) contended that if one type of eye movement is solely influenced by conscious memory, while another type is only influenced by unconscious memory, then this would provide strong evidence for the existence of distinct conscious and unconscious processes. To assess conscious memory, they focused on the comparison between old scenes receiving a 'recollect old' rating (6) and those receiving a 'sure old' rating (5); in contrast, to assess unconscious memory, they focused on the comparison between old and new scenes receiving a 'sure new' rating (1, the main focus of the sections above). Ramey et al. (2019) interpreted their FSA and scanpath efficiency findings as follows:

> Whereas conscious recollection for a scene uniquely improved the accuracy of the first eye movement in a search task, unconscious memory uniquely improved participants' search efficiency and gradually guided the eyes towards the target over the course of a trial… Furthermore, Bayesian analyses indicated that these memory effects on eye movements may be independent, such that conscious memory did not influence scanpath efficiency, and unconscious memory did not influence first saccade accuracy… (Ramey et al., 2019, p. 78).

In sum, Ramey et al. (2019) interpreted their findings as a whole as evidence of a double dissociation, namely that FSA was uniquely influenced by conscious memory whereas scanpath efficiency was uniquely influenced by unconscious memory.[5]

How compelling is this interpretation? Ramey et al. reported a significant difference in FSA between 'recollect old' and 'sure old' old scenes, but not between old and new scenes endorsed as 'sure new' (for which Bayesian evidence supported a null effect); in contrast, there was a significant difference in scanpath efficiency between 'sure new' old and new scenes, but not between old scenes receiving 'recollect old' and 'sure old' ratings (for which Bayesian evidence again supported a null effect). From this pattern it is clear that the dissociation rests on contrasts between statistically-significant and non-significant effects. But of course just because one effect is significant and another non-significant, it does not follow that the effects themselves are significantly different in magnitude (Nieuwenhuis, Forstmann, & Wagenmakers, 2011; Palfi & Dienes, 2020).[6] The reported Bayes factors in favor of the null effect in the two contrasts are also not diagnostic of whether conscious and unconscious memory uniquely affected FSA and SPR respectively.

To see this, Fig. 6A presents the raw effects of conscious and unconscious memory on FSA, SPR, and reaction time (RT), and their confidence intervals. A 2 (memory type: conscious vs. unconscious) x 2 (eye-movement measure: FSA vs. SPR) analysis of variance, after excluding an extreme outlier,[7] finds a non-significant interaction, $F(1, 14) = 3.23$, $p = .09$, $\eta^2 = 0.05$, and a Bayesian analysis confirms that the evidence is inconclusive ($BF_{10} = 0.95$). This is reflected in the overlapping confidence intervals of the conscious and unconscious effects on the two eye movement variables, such that the data only support the strong claim that "conscious recollection for a scene uniquely improved the accuracy of the first eye movement" to the extent that the conscious effect on FSA was statistically significant while the unconscious effect on FSA was not. But they challenge this claim to the more important extent that the conscious effect on FSA was not significantly larger than the unconscious effect on FSA. By the same token, the data only support the strong claim that "unconscious memory uniquely improved participants' search efficiency" to the extent that the unconscious effect on SPR was statistically significant while the conscious effect on SPR was not. But, again, they challenge this claim to the extent that the unconscious effect on SPR was not significantly larger than the conscious effect on SPR. It is invalid to interpret the significance of one effect and non-significance of another as evidence that the two effects are themselves significantly different.

While the raw effects in Fig. 6A allow us to assess Ramey et al.'s claims about the relative magnitudes of the conscious and unconscious effects on each eye-movement measure (and on RT), this may not be the most appropriate way to test whether a double dissociation is present in the data. The reason for this is that the conscious (difference in the relevant eye-movement measure between old scenes given a 'recollect old' versus a 'sure old' rating) and unconscious (difference between old and new scenes given a 'sure new' rating) effects are based on quite different contrasts, and there is little reason to think that these two types of difference score lie on a common measurement scale. To better characterize the influence of conscious and unconscious memory on FSA and SPR, Fig. 6B instead presents them in terms of effect sizes. This panel allows the double dissociation claim to be more directly assessed, with the conscious and unconscious effects being placed on a common scale across the two measures. When standardized in terms of Cohen's d_z, we see that the conscious effect on FSA was not significantly greater than that on SPR, and the unconscious effect on SPR was not significantly greater than that on FSA. This again demonstrates that the key double dissociation claim is not convincingly supported by the data. At most, the data reveal a (numerically) somewhat greater conscious effect on FSA than SPR and a somewhat greater unconscious effect on SPR than FSA.

It is worth noting that, in the above analyses, SPR-related estimates were obtained based on SPR data that included the first eye movement.[8] Since a significant effect of conscious memory on FSA was reported by Ramey et al. (2019), there may be concerns as to the validity of the comparisons presented above. Specifically, to the extent that FSA, a measure involving the first eye movement, was significantly affected by conscious memory, other measures involving the first eye movement, such as the aforementioned SPR estimates, may also be contaminated by this conscious effect. In other words, there may be shared variance between the SPR and the FSA estimates. Recognizing this potential issue, Ramey et al. (2019) included additional analyses in which only scanpath ratios from the second saccade onward were taken into account. These revealed the same pattern of results compared to results based on SPR including the first eye movement. Crucially, while the magnitude of the reported effect size of the unconscious effect on SPR was numerically smaller when the first saccade was excluded (first saccade excluded: $d = -0.24$; first saccade included: $d = -0.42$), the magnitude of the conscious effect on SPR was also numerically smaller when the first saccade was excluded (first saccade excluded: $d = -0.10$; first saccade

[5] Note also that Ramey et al.'s interpretation is clearly about memory processes and not just subjective experiences. The interpretation is about causes ('influence' is a causal term), yet subjective experiences do not have causal powers independent of the representations and processes on which they supervene.

[6] The experiment was sufficiently sensitive to detect an overall effect of scene repetition. Based on the original dataset, we found that new scenes differed significantly from old scenes, collapsed across all recognition confidence ratings, in terms of FSA, SPR, and reaction time (RT). For FSA ($M_{\text{new scenes}} = 87.75$ degrees error, $M_{\text{old scenes}} = 81.09$ degrees error), $t(22) = 2.44$, $p = .01$, $d_z = 0.52$; for SPR ($M_{\text{new scenes}} = 10.07$, $M_{\text{old scenes}} = 8.15$), $t(22) = 3.76$, $p < .001$, $d_z = 0.80$; for RT ($M_{\text{new scenes}} = 4617$ ms, $M_{\text{old scenes}} = 3973$ ms), $t(22) = 4.48$, $p < .001$, $d_z = 0.95$. These p-values are one-tailed.

[7] The extreme outlier was identified by the standard boxplot method, i.e., 3 interquartile ranges above the third quartile or below the first quartile.

[8] No scanpath ratio data based on the second saccade onwards are included in Ramey et al.'s (2019) dataset.

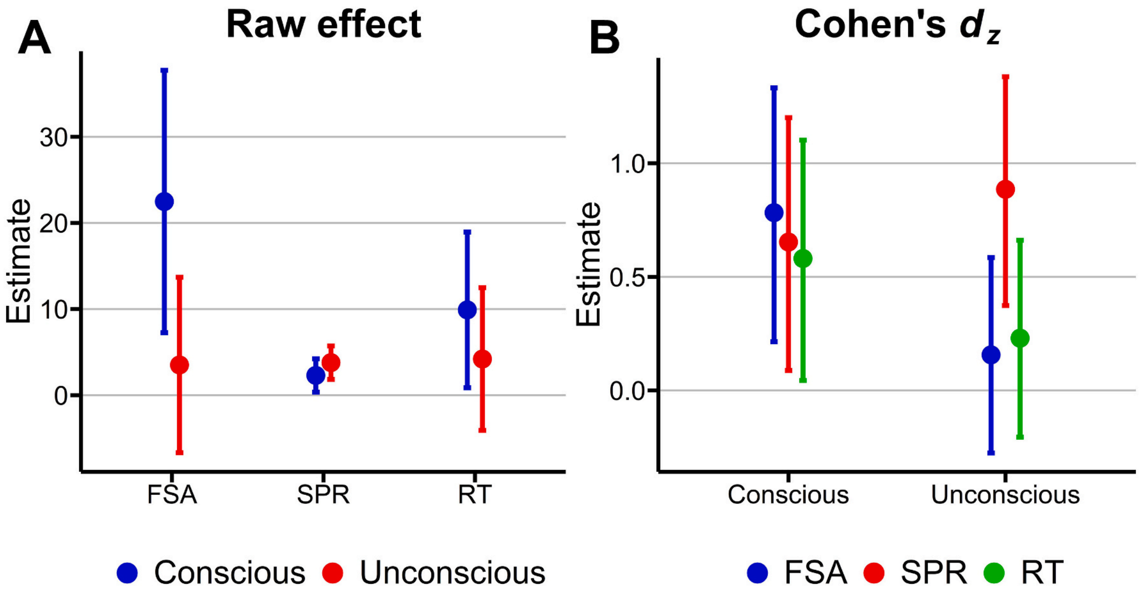


Fig. 6. (A) Raw effects of conscious and unconscious memory on first saccade accuracy (FSA), scanpath ratio (SPR), and reaction time (RT). RT was rescaled from milliseconds to deciseconds. (B) Effect sizes of conscious and unconscious influences on FSA, SPR, and RT. The error bars represent 95% confidence intervals (CIs). The CIs of the effect sizes were estimated using the noncentrality parameter method (Cumming & Finch, 2001).

included: $d = -0.17$). Indeed, it stands to reason that excluding the first eye movement would not only affect the SPR of both old and new items rated 1, but also the SPR of old items rated 5 and 6. Inasmuch as the interaction effect discussed above is concerned with whether the conscious and unconscious effects exhibited different slopes for FSA and SPR measures, if the effect size estimates for both conscious and unconscious effects on SPR are reduced due to exclusion of the first eye movement, then the interaction effect is likely to remain non-significant. Assuming, for the sake of argument, that excluding the first eye movement entails a greater reduction in the magnitude of the estimated effect size of unconscious memory on SPR (as suggested by results reported by Ramey et al., at least numerically), this will result in a shallower slope between conscious and unconscious effects for SPR. In other words, a significant interaction effect is even less likely to be revealed.

Finally, we assessed whether the required pattern of conscious and unconscious effects was manifested at the individual-participant level. After excluding 7 participants with missing data in one of the 4 conditions (2 memory types: conscious vs. unconscious; 2 eye-movement measures: FSA vs. SPR), as well as one extreme outlier, we found that only 7 out of the remaining 15 participants exhibited a dissociation pattern, namely a larger conscious than unconscious effect on FSA as well as a larger unconscious than conscious effect on SPR (See Appendix A.3 for more information on participant-specific patterns). All in all, the data lend little weight to Ramey et al.'s (2019) claim of double dissociation.

7. Experiments 1 and 2

Insofar as our simulation results indicate that the regression to the mean phenomenon may arise from *post hoc* data selection, the single-system model makes a novel and testable prediction in a two-stage awareness test. Suppose that Ramey et al.'s (2019) procedure is repeated and that in the test stage (Fig. 1B), old and new scenes receiving ratings of 1 are identified. Although these items receive the same recognition confidence rating, the model hypothesizes that they do not have the same true, latent memory strength: S is larger for old items, but is compensated for by a smaller (more negative) value of e_{REC}. It follows that if we now administer a second awareness test (Fig. 1C and D), focusing on the old and new scenes rated 1 in the familiarity judgment stage, participants should reliably differentiate between them. If we assume that the errors are independent each time the latent strength of an item is sampled from memory, then e_{REC} will be identical for the selected old and new items, but the old items will continue to have larger S values.

To test this prediction, we conducted two online studies in which an experimental design similar to Ramey et al.'s (2019) was adopted with a second awareness test. Specifically, the second awareness test comprised either a two-alternative forced choice (2AFC) task (Experiment 1; Fig. 1C) or a single-item recognition rating task (i.e., the same recognition measure adopted in the familiarity judgment stage) (Experiment 2; Fig. 1D). Due to the online nature of the experiments, no eyetracking measure was incorporated; instead, akin to conventional contextual cueing experiments, RT across trials was taken as the primary indicator of participants' search performance.

We expected that participants would exhibit overall enhanced search performance in the initial test stage for images that were repeatedly presented (i.e., old images) compared to non-repeated images (i.e., new images) and, additionally, that they would exhibit improved search RT for old images receiving the lowest familiarity confidence rating compared to new images receiving the same rating. Furthermore, following our simulation results, we predicted that search RT would be only weakly associated with familiarity, such that old images receiving lower recognition confidence ratings would be associated with slightly higher search RT. More crucially, regarding the 2AFC task (Experiment 1), we expected that participants would demonstrate higher-than-chance accuracy in identifying old images that had been judged new in the preceding part of the test phase, whereas with respect to the second single-item recognition rating task (Experiment 2), we expected that old scenes would be rated higher than new scenes. These two experiments were preregistered on the Open Science Framework (respectively https://osf.io/ht9z3 and https://osf.io/246ha) and the data are available at https://osf.io/8nfqj/ and https://osf.io/eyrhd/ respectively.

7.1. Methods

7.1.1. Participants

For Experiment 1, in which the 2AFC task was adopted, an *a priori* analysis using G*Power (version 3.1.9.7; Faul, Erdfelder, Lang, & Buchner, 2007), conducted before data collection, indicated that a sample of 93 participants would be needed to detect a correlation between search performance and familiarity ($r = -0.255$) with a power of 0.80 at a one-tailed alpha level of 0.05. A total of 107 participants located in the UK (48 males and one non-binary; $M_{age} = 36.30$, $SD_{age} = 10.39$, range = 18–59) were recruited via Prolific. For Experiment 2, in which a second single-item recognition rating task was added to the standard task, another *a priori* analysis using G*Power and conducted before data collection indicated that a sample of 71 participants would be needed to detect a small effect ($d_z = 0.30$) in the second awareness test with a power of 0.80 at a one-tailed alpha level of 0.05. A total of 90 participants located in the UK (40 males and one non-binary; $M_{age} = 39.78$, $SD_{age} = 10.77$, range = 20–60) were recruited via Prolific. In both experiments, all participants had normal or corrected-to-normal vision and were asked to complete the experiment via a web browser in a quiet environment without distractions. Informed consent was obtained from all participants. Participants in the first experiment were paid £3.75, whereas those in the second experiment were paid £6 (due to a longer experimental procedure and inflation). Both studies were approved by the UCL Research Ethics Committee.

7.1.2. Materials

The experiments were programmed with PsychoPy (Peirce et al., 2019). A total of 96 realistic full colored images of indoor and outdoor scenes were used as stimuli. The images were displayed at an aspect ratio of 4:3 irrespective of the device on which the experiment was run. In Experiment 1, a small red letter ‘T’ or ‘L’ with a letter height of 0.03 serving as the search target was embedded in each image during the first two stages. After excluding the center (0.1×0.1 in *height* units) and the periphery (outer 15%) regions of the images as the spawning locations for the search targets, the exact locations of the targets were randomly generated such that they were evenly spread over the four quadrants of the screen across the images. The identity of the target letter for each scene was also predetermined from the outset and remained constant throughout the experiment. A counterbalancing measure was adopted to reduce potential stimulus effects. Specifically, the 96 images were separated into three equal sets with 32 images each. One set of images was then randomly assigned as lures to be introduced in the familiarity judgment stage, while the other two sets were assigned as stimuli for the search stage, across participants.

In Experiment 2, two slight changes were made. The size of the target letters was set at a smaller letter height of 0.015, so as to reduce the possibility of a ceiling effect on participants' search speed; the contextual cueing effect was expected to be increased as a result of this change. In addition, an improved counterbalancing method was adopted. Specifically, at the beginning of the experiment, a set of 64 images was randomly selected from the overall pool of the 96 images to be used as the old scenes across the participants, while the remaining 32 images were used as lures (i.e., new scenes).

7.2. Procedure

Both of the experiments, consisting of three stages, were implemented on the Pavlovia website (pavlovia.org). Fig. 1 illustrates the tasks. For both experiments, the first two stages were the search stage (Fig. 1A) and the familiarity judgment stage (Fig. 1B) respectively. In these two stages, Ramey et al.'s (2019) procedure was followed closely. The experiments could be run on any device of the participant's choice, apart from mobile devices. In the search stage, participants completed a learning phase of 64 trials during which they looked for target letters across a series of randomly ordered images. Each trial started with a fixation cross presented at the center of the screen for 1.5 s followed immediately by presentation of the image. In each trial, participants were asked to locate the search target in the image and respond to the identity of the target letter using either the key <T> or <L>. Each trial ended when a response was registered, or after 20 s in the absence of a response. If the response was incorrect, or no response was given within 20 s, an error message in red was presented at the center of the screen for 1 s before moving on to the next trial. No indication of a subsequent memory test phase was provided. Participants were given a short break upon completion of the first stage before proceeding to the test phase.

The familiarity judgment stage of the experiments followed the search stage. After presentation of a fixation cross at the beginning of each trial, a preview of an image lasting 0.4 s without its target was presented. The image was either one that had appeared in the search stage or a new one. Immediately after each preview, participants were instructed to report whether or not they thought the image had been presented in the search stage, with no time limit. Responses were recorded based on a combination of a recollection response option with a value 6 (*Recollect old*), and a 5-point familiarity confidence scale, with values 5 (*I'm sure it's old*), 4 (*Maybe it's old*), 3 (*I don't know*), 2 (*Maybe it's new*), and 1 (*I'm sure it's new*). An explanation of the rating options was provided before the commencement of the second stage. Specifically, participants were informed that a response of ‘recollect old’ (6) should be selected if they could recall details of their experience of having seen the image in the first stage, such as an emotion felt or sensations experienced when they viewed the image previously. In contrast, participants were instructed to select a response of ‘sure old’ (5) if they were certain that they had viewed the image while not being able to recall any accompanying episodic details. After giving the rating response in each trial, another fixation cross at the center was presented to participants, followed by presentation of the same scene, though with the search target this time. Participants were instructed to search for the target letter as they did in the first stage. There were 96 trials (64 old images and 32 new images) in the familiarity judgment stage, and the images were randomly ordered as before. Participants were given another short break upon completion of this stage.

For Experiment 1, the third and final stage was the 2AFC stage (Fig. 1C). Before the 2AFC task began, participants received the following instructions: “In the final stage of the experiment, two images will be shown side-by-side on the screen in each trial. One of the images will be one that you have viewed twice since the beginning of the experiment, while the other image will be one that was only presented to you once during the second stage of the experiment. Your task is to choose the image that you have viewed twice by pressing either the ‘left’ or ‘right’ arrow key. You may take as long as you need to think carefully and make a decision in each trial.” Each trial began with a fixation cross, after which a pair comprising an old and a new image were presented side-by-side, with another instruction, serving as a reminder, being shown near the top of the screen. This stated that participants should select the image they had viewed twice since the beginning of the experiment by pressing either the ‘left’ or ‘right’ arrow key. Participants were given as long as they needed to select the old image (i.e., the image that had appeared twice previously), and the trial ended only when they made a decision by pressing one of the arrow keys.

A programming snippet was written to ensure that, whenever possible, each image pair comprised a new and an old image receiving the same recognition rating in the second stage. Specifically, the pairing algorithm took each new image given a rating of 1 in the first stage of the test and paired it with an old image receiving the same rating, then did the same for new images given a rating of 2, and so on. In cases where there were new images left unpaired because there were not enough old images given the corresponding rating, these new images were paired with randomly chosen old images given different ratings (these pairings were excluded from the key analysis, of course). As a result of the pairing process, the 2AFC task comprised a total of 32 trials. The sequence of image pairs appearing in this task was again randomized.

In contrast, for Experiment 2, the final stage (Fig. 1D) involved a second single-item recognition rating task. Before the task began, participants received the following instructions: “In the final part of the experiment, you will view the images from the previous stages. Your task is to provide a memory response indicating whether you recognize the image from the first part of the experiment by pressing the corresponding number key above the top row of letters. You may take as long as you need to make a judgment." Each trial began with a fixation cross, after which an image was presented, accompanied by a question and the rating options shown at the bottom of the screen. The question asked whether the image was from the first part of the experiment. Participants were given the same 6-point rating options as in the familiarity judgment stage and were given as long as needed to arrive at a decision. Unlike the task in the familiarity judgment stage, the trial ended as soon as the participant selected a rating response, and no search task was incorporated into this stage. There were 96 trials (64 old images and 32 new images) in this stage, and the presentation sequence of the images was again randomized.

After completion of the third stage in both experiments, participants were thanked and debriefed.

7.3. Data pre-processing and analysis plan

With respect to Experiment 1, following the pre-registration, in order to ensure that only those participants who were sufficiently attentive to the experimental tasks were included in the subsequent analyses, we excluded 13 participants who failed to achieve 95% accuracy in reporting the correct identity of the search target in the first two stages. Thus, a final sample of 94 participants from this study was included in the subsequent analyses. For Experiment 2 the exclusion threshold was lowered to 85% accuracy, as stated in the pre-registration, since the reduced size of the target letters was expected to increase the difficulty of the search task. This resulted in exclusion of 7 participants and a final sample of 83 participants.

Because the two experiments were very similar across the first two stages, the data were pooled for statistical analyses. Trials with incorrect search responses were excluded (2.96% of total trials). Statistical tests were conducted separately for the third stage of the two experiments. We set an *a priori* exclusion criterion for Experiment 2 such that data in the final recognition stage for participants who gave a rating of 1 to fewer than 3 old scenes or 3 new scenes in the familiarity judgment stage would be excluded. However, no equivalent criterion was set for Experiment 1, as the number of trials in the 2AFC stage was already comparatively small. As already mentioned, *a priori* power analyses were conducted for the experiments based on a one-tailed alpha level of 0.05. Hence, *p*-values were estimated based on one-tailed tests for the *t*-statistics reported below, where we had strong *a priori* predictions in line with our hypotheses. It is also noteworthy that in the pre-registration of Experiment 1, we originally planned to exclude trials with RTs longer than 10 s before conducting the analyses in relation to the familiarity judgment stage. This exclusion threshold had been adopted in previous contextual cueing studies conducted in our laboratory. However, in light of the fact that no such threshold was adopted by Ramey et al. (2019) and that in our previous studies no search tasks involving real world images were incorporated, we conducted the main analyses below without applying the exclusion threshold. We further conducted reliability analyses on the critical measures in the familiarity judgment stage and the final test stages of the two experiments, which can be found in Appendix A.4. In addition, potential issues relating to outliers were addressed in analyses in Appendix A.5. All the statistical analyses were carried out in R (R Core Team, 2022).

7.4. Results

7.4.1. Contextual cueing effect

First, to investigate whether a contextual cueing effect, comprising facilitated search performance due to repeated exposure to the stimuli, emerged across Experiments 1 and 2 combined, we focused on the comparison of participants' search performance (in terms of RT) for old and new images across all recognition confidence ratings during the familiarity judgment stage. A paired-samples *t*-test indicated that there was a significant difference between their search speed for new ($M = 2.89$ s, $SD = 1.02$ s) versus old images ($M = 2.69$ s, $SD = 0.94$ s), $t(176) = 4.87$, $p < .001$, $d_z = 0.37$, suggesting that an overall contextual cueing effect was present. (See Appendix A.6 for an exploratory analysis of the cueing effect at the item level.)

7.4.2. Performance in the familiarity judgment stage

Recognition confidence ratings. The percentages of old scenes receiving ratings of 1–6 were, respectively, 16.81%, 20.97%, 16.60%, 16.18%, 14.75%, and 14.69%; in contrast, for new scenes, the percentages were, respectively, 35.32%, 28.04%, 15.41%, 11.41%, 6.82%, and 3.00%. These distributions suggest that participants exhibited a higher level of confidence when recognizing old images compared to new images, as expected.

Search speed. Fig. 7 illustrates search speed in the second stage, in terms of RT, as a function of memory response rating and image type.[9] Following our simulation results, our primary focus here was on whether there was any difference in search performance between old and new images receiving a rating of 1 (i.e., 'sure new'). A secondary question is whether search performance for old images was associated with familiarity confidence (i.e., ratings of 1–5). Regarding the former, mean search RT for old images which participants endorsed as 'sure new' was significantly lower than for new images receiving the same rating, $t(162) = 4.79$, $p < .001$, $d_z = 0.38$, indicating that the participants were quicker at finding targets in old images compared to new images even though they failed to recognize the old images. This represents a conceptual replication of Ramey et al.'s key SPR result but with search RT as the dependent measure.

Mirroring the statistical strategy adopted by Ramey et al. (2019), we assessed the relationship between search performance for old images and familiarity confidence using a linear mixed effects model (LMM) with crossed random effects of image and participant so as to adjust for potential influence of stimulus effects and individual differences.[10] We analyzed the data at the trial level and fitted the LMM using restricted maximum likelihood estimation based on the 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015) while calculating the degrees of freedom of the predictors using Satterthwaite approximation based on the 'lmerTest' package in R (Kuznetsova, Brockhoff, & Christensen, 2017). The results (see Table 2) indicated that search RT for old images was not significantly predicted by familiarity confidence (ratings 1–5), $b = 0.01$, $t(9328) = 0.79$, $p = .21$. This result conceptually replicates the absence of association already discussed above (see Fig. 2), and moreover is in congruence with Ramey et al.'s (2019) data: in a re-assessment of their data based on the same LMM specifications, we found that their participants' search RT for old scenes was not significantly associated with familiarity strength, $b = -0.05$, $t(1168.55) = -0.77$, $p = .22$.

In an exploratory analysis we further investigated whether the non-significant association between search RT for old images and familiarity confidence was consistent across the two experiments. Specifically, we included experiment (i.e., Experiment 1 vs Experiment 2) and the interaction term between experiment and familiarity as two additional predictors in the LMM model. The results indicated that the main effect of familiarity on search RT for old images remained non-significant, $b = -0.01$, $t(9327) = -0.33$, $p = .37$, whereas the main effect of experiment was significant, $b = 0.89$, $t(388.1) = 5.83$, $p < .001$, which can be attributed to the longer RTs due to the increased task difficulty (i.e., smaller target letters) in Experiment 2. More crucially, the interaction effect of experiment and familiarity was non-significant, $b = 0.05$,

[9] Fig. 7 was plotted based on the linear mixed model as described below. All the estimates were derived from the LMM using the 'emmeans' package in R.

[10] The model specification that we entered in the lmer function was: Response time ~ familiarity confidence (ratings 1–5) + (1|*subject*) + (1|*image*). The random effects are italicized. Including random slopes in the LMM models resulted in singular fits. Thus, we decided to treat familiarity as a fixed predictor not subject to any random effect.

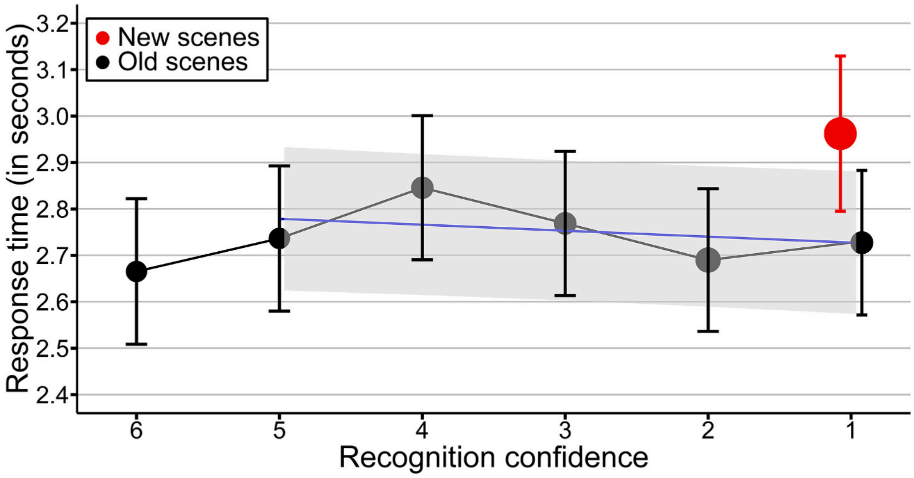


Fig. 7. Estimated marginal response time (RT) means across different recognition confidence ratings during the familiarity judgment stage in Experiments 1 and 2 combined. Symbol sizes represent the relative proportions of scenes across different recognition confidence ratings. The error bars represent the standard errors of the RT means. The blue line represents the regression line and the grey band represents the associated standard errors of the regression line across ratings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Indices of the linear mixed model assessing association between RT and familiarity ratings.

	Response Time		
Predictors	Estimates	CI	p
(Intercept)	2.71	2.41 – 3.02	**< 0.001**
Confidence rating	0.01	−0.02 – 0.04	0.21
Random Effects			
σ^2	3.67		
$\tau_{00\ \mathrm{subject}}$	0.87		
$\tau_{00\ \mathrm{image}}$	1.67		
ICC	0.41		
N_{image}	96		
N_{subject}	177		

Observations 9400
Marginal R^2 / Conditional R^2 0.000 / 0.409

t(9302) = 1.53, p = .13, suggesting that there was no difference in the relationship between search RT and familiarity across the two experiments.

7.4.3. Performance in the 2AFC stage in Experiment 1

Mean accuracy for selecting the correct image, across pairs of images receiving the same ratings in the familiarity judgment stage, was the main dependent variable for the 2AFC task in this stage. Specifically, mean accuracy was represented in terms of the percentage of trials in which participants correctly selected the old image. As participants encountered a pair of old and new images in each trial, chance performance is 50%. Our primary focus was on the pairs of images receiving ratings of 1, as these pairs incorporated the most confident misses and correct rejections (i.e., old and new scenes rated 1) in the previous stage. Among these pairs, participants exhibited above-chance accuracy (M = 55.00%, SD = 24.62%), t(85) = 1.88, p = .03, d_z = 0.20. In a further exploratory investigation, we excluded 9 participants who encountered fewer than 3 pairs of images rated 1 in the 2AFC task before repeating the same analysis. This approach brings the exclusion criterion of the final test stage data in line with that of Experiment 2. Again, above-chance accuracy was revealed, (M = 58.18%, SD = 19.37%), t(76) = 3.71, one-tailed p < .001, d_z = 0.43. Fig. 8A illustrates the respective distributions of mean accuracy across participants. (See also Appendix A.7 for an exploratory analysis of the number of trials encountered by participants across different accuracies for image pairs rated 1.) Together, these results indicate that participants on average exhibited

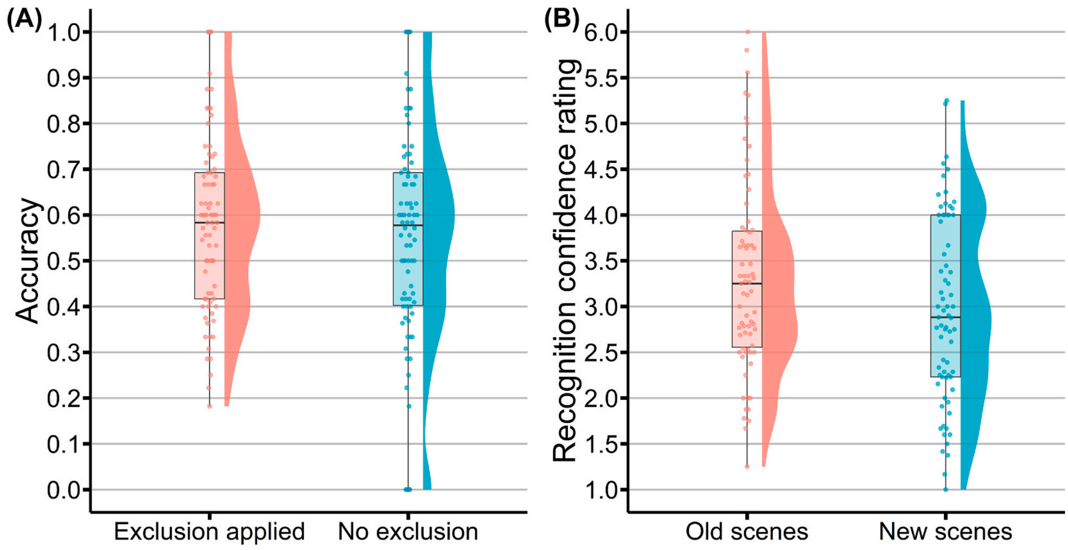


Fig. 8. (A) Distributions of mean accuracy in the 2AFC task of Experiment 1 for pairs of images previously rated 1 (*'sure new'*). The data are separated depending on whether the exclusion criterion of Experiment 2 was applied or not. (B) Distributions of mean ratings for old and new images in the final recognition stage of Experiment 2, limited to images that had been rated 1 in the familiarity judgment stage.

signs of awareness, as they were able to differentiate old images from new ones, despite having judged the old images as new with the highest level of confidence during the familiarity judgment stage.

Furthermore, by examining the 2AFC data at the participant level, 29 participants with chance level or worse accuracy, and receiving more than 3 trials rated 1, were identified. One interesting question is whether these participants would nonetheless exhibit a significant cueing effect among images rated 1 in the familiarity judgment stage. This is because these participants not only judged these old images as new with highest confidence in the familiarity judgment stage, but also failed to differentiate them from new images rated 1 in the 2AFC stage, suggesting that they genuinely did not recognize such old images. We investigated this question in an exploratory analysis. No significant RT difference between old and new images rated 1 in the familiarity judgment stage was revealed, $t(28) = 0.49$, $p = .31$, $d_z = 0.09$, and a Bayesian analysis also provided moderate support for the null difference ($BF_{10} = 0.30$). Thus there was no evidence of contextual cueing in participants whose 2AFC recognition memory was at or below chance.

7.4.4. Performance in the second recognition rating stage in Experiment 2

Regarding the final recognition stage of Experiment 2, the main focus here was on whether there was any difference in recognition confidence ratings between old and new images which had received a rating of 1 in the preceding familiarity judgment stage. Among the 71 participants who gave at least 3 old images and 3 new images ratings of 1 in the previous stage, a significant difference in rating between such old ($M = 3.30$, $SD = 1.05$) and new images ($M = 2.97$, $SD = 1.02$) was revealed in this final recognition stage, $t(70) = 3.82$, $p < .001$, $d_z = 0.46$ (see Fig. 8B). This result again suggests that participants on average were able to differentiate between old and new images that they previously judged as new with the highest confidence, in congruence with the 2AFC results.

Furthermore, as there were 22 participants who gave an equal or higher mean rating to new images compared to old images in this stage, another exploratory analysis focusing on these participants was conducted, resembling that reported above for the 2AFC data. The fact that these participants did not on average rate old images higher than new images in the final recognition stage suggests that they were unable to differentiate them across the two recognition tests. Hence, it is interesting to investigate whether they nonetheless exhibited a cueing effect in relation to images rated 1 in the familiarity judgment stage. The result indicated that there was no cueing effect among these participants, $t(21) = 1.24$, $p = .12$, $d_z = 0.27$, $BF_{10} = 0.76$.

7.5. Discussion

Regarding Experiment 1, our prediction based on the single-system account is tested and confirmed by the results of the 2AFC task, where participants showed higher-than-chance accuracy in identifying old versus new images that they judged as new with the highest level of confidence in the previous stage. Although these items received the same recognition confidence rating in the familiarity test, the model hypothesizes that they do not have the same true, latent memory strength: S is larger for old items, but is compensated for by a smaller (more negative) value of e_{REC}. Thus, in the subsequent 2AFC test, the model predicts that old items rated 1 will be reliably selected (assuming that e_{REC} no longer differs because of independent sampling). Had the cueing effect for old images endorsed as new been brought about by unconscious learning, participants would have demonstrated chance level performance in identifying these very same images in the 2AFC task.

Could this contrast between the single-item and 2AFC test stages be reconciled with Ramey et al.'s (2019) theoretical perspective? It would not help to propose, as some have (e.g., Bastin & Van der Linden, 2003; but see Bayley, Wixted, Hopkins, & Squire, 2008; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000), that 2AFC may be more sensitive than single-item recognition to conscious memory signals such as familiarity, because the whole point of contrasting highest confidence misses and correct rejections is to equate them for conscious memory. Hence such a conjecture would be tantamount to conceding that the familiarity judgment task in Ramey et al.'s experimental method is insensitive to subtle differences in conscious memory. On the other hand, if 2AFC is more sensitive than single-item recognition to *unconscious* memory signals, then the results of this experiment could be accommodated within Ramey et al.'s framework. Such a view would propose that different patterns of unconscious eye movements elicited by the old and new images within each 2AFC pair can be used as the basis for a recognition decision, while at the same time the unconscious eye movements elicited by the same old and new images cannot support recognition decisions when the items are presented individually. Such a strong assumption would require independent verification, but in any case is manifestly less parsimonious than the single-system explanation.[11]

In any event such speculation is rendered moot by the results of the final recognition stage of Experiment 2 in which old and new items previously rated '*sure new*' were tested individually. The significant difference in ratings in this novel test lends even stronger support to our prediction and the regression to the mean hypothesis, since the recognition measure adopted in this stage was identical to the one in the familiarity judgment stage. No explanation based on differential sensitivity can be made in this context. To the extent that participants on average rated old images higher than new images, both of which had been rated 1 previously, the result provides compelling evidence in favor of our hypothesis based on the single-system model and challenge a fundamental assumption of Ramey et al.'s (2019) logic, namely that old and new items rated new with high confidence in the familiarity test are equated for conscious memory.

Remarkably, when focusing on data of participants who did not on average give a higher rating to old compared to new images in the final test, no significant RT difference between old and new images rated 1 in the familiarity judgment stage was revealed, suggesting that the cueing effect vanished among these participants. The same pattern was found for Experiment 1. This again is in line with our model: assuming respective e_{REC} values associated with these old and new images are not different in this stage because of independent sampling, this would suggest that the latent memory strength S is not larger for old images among these participants. In this context, the model predicts that no facilitation in RT among these old images would occur. This is exactly what was observed in the exploratory analysis.

It is noteworthy that, as the target identity was pre-determined and fixed for each image from the outset, there could be potential concerns in relation to the reported results involving RT as the dependent variable. This is because, unlike conventional array-based contextual cueing tasks, there is a possibility that participants could develop a strong association between the target identity and the scene in the learning stage. The observed cueing effect could thus be observed even if no search took place, and the RT difference may not reflect facilitated search performance. To address this potential issue, we re-assessed Ramey et al.'s (2019) eyetracking data. Specifically, restricted to old scenes presented once in the test phase, we focused on the distance between the participant's fixation and the target at the last saccade in each trial of the search task. It was revealed that the mean distance between the participant's fixation and the target was 29.95 pixels,[12] and the maximum distance was below 50 pixels. This suggests that all participants across

[11] In the case of other indices of unconscious memory, there is clear evidence against this view: the densely amnesic individual E.P. shows normal priming – on this perspective, an unconscious memory signal – combined with chance-level 2AFC recognition (Hamann & Squire, 1997).

[12] While no information was provided regarding the unit of measurement for distance in the shared data, Ramey et al. (2019) referred to pixels when describing how the stimuli were prepared in their study.

all trials completed the search for the target before responding. Therefore, we suggest that the issue regarding potentially strong associations between the target identity and the scene may be overstated.

In sum, these findings, together with our simulation results, suggest that the seemingly unconscious cueing effect for old images judged as new in the familiarity judgment stage is rooted not in truly unconscious learning, but in statistical artifacts due to the inherent flaws of the *post hoc* data section approach.

8. General discussion

One widely employed method to demonstrate unconscious learning is to select *post hoc* items for which participants appear to lack some aspect of awareness (in this case, awareness of repetition as measured by familiarity), and check if performance is nonetheless enhanced by that aspect. By means of this method, Ramey et al. (2019) showed that old scenes endorsed as new with high confidence were associated with a lower average scanpath ratio than new scenes receiving the same recognition response. Ramey et al. attributed the enhanced scanpath efficiency to unconscious learning. Nevertheless, as demonstrated in this article through a *post hoc* model fitting simulation and a novel experimental test, their results can be satisfactorily accounted for by a simple single-system model that assumes no unconscious memory representation at the latent level. Therefore, the scanpath ratio findings cannot be taken as unequivocal evidence of unconscious learning.

The model suggests that Ramey et al.'s key finding is a statistical inevitability that stems from a ubiquitous but regularly overlooked phenomenon—regression to the mean, arising from *post hoc* data selection. While it is perfectly reasonable to assume that errors are randomly distributed on average, as soon as the researcher selects items on the basis of their observed values, this randomness no longer holds. Part of the very reason why a scene is given an extreme recognition confidence rating is because its error component happens to be extreme. Independence between error and true score does not entail independence between error and measured score. Regression to the mean then comes into play when a second measure is taken (*SPR* in this case) in which the error is no longer biased by selection. Fundamentally, we submit that the *post hoc* selection method – as instantiated here via the selection of high-confidence misses and correct rejections – is intrinsically not an appropriate tool to demonstrate unconscious processes (see also Rothkirch et al., 2022; Shanks, 2017).

The novel prediction tested in two preregistered experiments, in both of which an additional awareness test was incorporated, lends support to this interpretation. Specifically, as shown in Experiment 1, even when old and new images received identical 'sure new' ratings (1) in the familiarity stage, participants were able to differentiate the old from new images in the subsequent 2AFC test. In addition, corroborating evidence is provided by the results of Experiment 2, in which the final awareness test adopted the same recognition confidence measure as used in the familiarity stage. Focusing on old and new images which were judged as new with the highest level of confidence in the familiarity judgment stage, participants nonetheless were able to differentiate between these two types of images by giving old images significantly higher recognition ratings in the final awareness test. Indeed, even setting aside our simulation results, the experiments reported here are sufficient to critically undermine the logic of Ramey et al.'s approach. Their analysis presupposes that conscious memory is equivalent between old and new scenes endorsed as 'sure new' and that therefore the lower average scanpath ratio for old scenes must result from unconscious memory. But our experiments show that old and new scenes judged 'sure new' in the familiarity test actually do not have equal latent memory strengths. This becomes evident in the subsequent recognition test in both experiments in which old scenes were reliably differentiated from new ones. On this basis, it can no longer be argued that the lower average SPR for old scenes is a hallmark of unconscious memory.

In addition to explaining the SPR effect for unrecognized scenes, the null association between familiarity and scanpath ratio that Ramey et al. (2019) observed is also reproduced by the model. It is striking that a model which predicts above-chance recognition and contextual cuing on the basis of a common latent representation can at the same time predict a correlation of effectively zero between these measures. Many other researchers have taken a statistically nonsignificant correlation between an implicit and an explicit measure as evidence of unconscious processing. Even though Ramey et al. reported additional results based on Bayesian analyses, which indicated moderate evidence in favor of a null association, this must be interpreted with caution, as such evidence is weakened when measurement uncertainty is taken into consideration (Malejka et al., 2021), particularly when weak correlations are expected to be more likely *a priori*. In addition to reviewing the problems with the approach which include wrongly slipping from a failure to reject the null to acceptance of the null, and demonstrating the danger of not accounting for uncertainty when conducting Bayesian analyses, Malejka et al. (2021) describe alternative Bayesian methods. These permit Bayes factors with scientifically informed priors to be calculated which reflect evidence for the null. Malejka et al. (2021) also provided formal guidance on how to conduct sample size planning. This allows the minimum sample size to be estimated for the given reliability of the two measures that would yield clear evidence in favor of a null correlation, if true.

Ramey et al. (2019) interpreted their findings of a significant conscious and non-significant unconscious effect on FSA, in conjunction with a significant unconscious and non-significant conscious effect on scanpath efficiency, as evidence of a double dissociation. This interpretation rests, however, on the fallacy of interpreting the significance of one effect and non-significance of another as evidence that the two effects are themselves significantly different. Again, even though Ramey et al. (2019) reported Bayes factors in favor of the null for the unconscious effect on FSA and that for conscious effect for scanpath ratio, such results do not demonstrate that conscious and unconscious effects affected the two eyetracking measures differentially. In this regard, Fig. 6 shows that in fact there was no statistically significant difference between the conscious and unconscious effects on either FSA or SPR, and the interaction was not significant, where the Bayesian analysis provided no compelling evidence to suggest otherwise. Hence, Ramey et al.'s (2019) claim of double dissociation of conscious and unconscious influences on the two eye-movement measures is not supported by the data.

Our claim is not that all aspects of eye movements are under conscious control. There are highly likely to be cues (e.g., luminance or motion) that influence eye movements independently of awareness (Spering & Carrasco, 2015). Rather our claim is that awareness is a necessary condition for *memory-guided* influences on eye movements. To the extent that learned statistical regularities, such as between particular scenes and target locations embedded within them, can affect eye movements, we conjecture that these regularities are available to conscious reports. Similar claims have been defended in relation to other seemingly 'automatic' behaviors such as Pavlovian conditioned responses (Lovibond & Shanks, 2002).

Although the pattern of eyetracking results may be accounted for by the single-system model, admittedly this alone does not necessarily rule out the possibility of unconscious learning. The model could be falsified – and unconscious learning demonstrated – in a number of ways, for example if no discrimination of high-confidence misses and correct rejections occurs in a second memory test or if reliable contextual cuing occurred in participants whose discrimination is at chance in the second test. An important direction for future research, thus, is to construct alternative multiple-system models that assume distinct systems for conscious and unconscious memory, and pitch them against the current single-system model. If they offer better model fits for eyetracking data based on standard model-selection criteria, such a finding may potentially serve as evidence supporting the multiple-system perspective (see Berry et al., 2012, for an example of this approach). Furthermore, in this work we have not considered individual differences, as our main focus has been on the interpretation of Ramey et al.'s (2019) key qualitative results for unconscious processing. There may be differences in, for instance, the magnitude of the correlation between familiarity and scanpath ratio across participants. In future work the modelling could be extended to address such issues.

In summary, the analysis presented here is part of a tradition going back at least 20 years (e.g., Shanks & Perruchet, 2002) in which artifacts arising from regression to the mean, appearing to reveal unconscious influences, can instead be shown to be consistent with single-system models. We submit that the model described here offers a cogent and parsimonious account of the memory-based control of eye movements in context-guided visual search.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data availability

We have shared the links to our data/code in the main text.

Acknowledgments

This research was supported by a grant from the United Kingdom Economic and Social Research Council (ES/S014616/1). We thank Simone Malejka, Miguel Vadillo, David Luque, and Christopher Berry for their helpful suggestions, and Michelle Ramey for her valuable feedback on an early draft of this paper. Code for the simulation reported here, and data from the experiments, are openly available at https://osf.io/8nfqj/ and https://osf.io/eyrhd/.

Appendix A

A.1. Breakdown of the S and e_{SPR} values across different SPR bins

To further verify that, irrespective of which variable is selected *post hoc*, the single-system model makes the same prediction about regression to the mean, we assessed the mean values of S and e_{SPR} for old and new scenes across different SPR bins based on the simulated data. As illustrated in Fig. A.1, a pattern of results similar to that shown in Fig. 4 emerged.

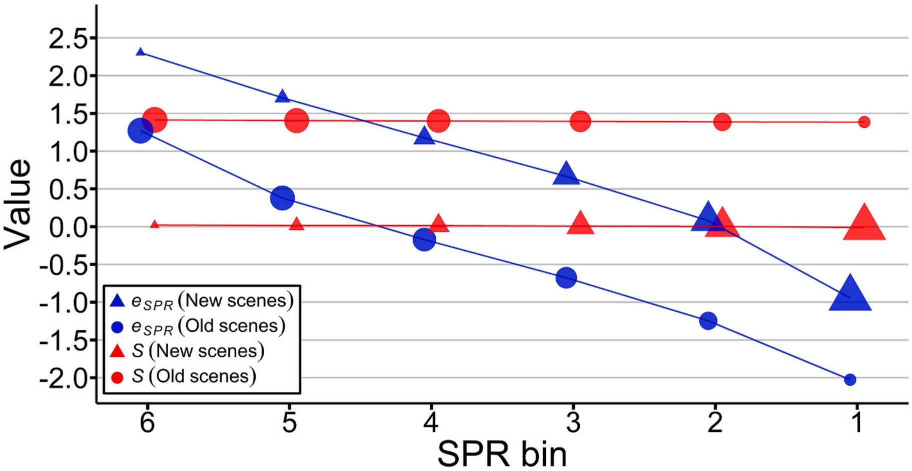


Fig. A.1. Mean values of S and e_{SPR} for old and new scenes across different SPR bins in the model simulation. Symbol sizes represent relative proportions of scenes across different SPR bins.

A.2. An extended model including between-participants variance

Here we present an extended version of the model which incorporates between-participants variance. All aspects are identical to the model described in Section 2.1 except for the following.

We first sample a mean value of S (old) for each participant from a normal distribution with a mean of 1.4 and SD of 0.5. Then, trial-by-trial S (old) values are sampled from each participant's distribution with the respective mean S (old) (and $SD = 0.1$). In addition, instead of binning each participant's REC values into 6 recognition ratings based on the overall proportions observed in Ramey et al.'s (2019) data, we aggregate all REC values collapsed across participants, and then divide them into 6 recognition bins in accordance with the overall proportions. This approach ensures between-participants variance in recognition confidence ratings, as different participants now generate different sets of proportions of responses across the 6 recognition bins, unlike the simulated data reported in the main text. Lastly, to approximate the observed reliability for the SPR measure, the SD of e_{SPR} is slightly decreased from 1 to 0.8.

As illustrated in Fig. A.2, the results generated by this simulation are virtually identical to those shown in Fig. 2. In addition, the simulated data for the novel test (i.e., reversing the SPR and recognition confidence variables) after incorporating between-participants variance are also similar to the results reported in Section 4. The simulated Spearman-Brown corrected reliability estimates based on 5000 random splits for the recognition confidence and SPR measures are $r = 0.88$ and $r = 0.68$ respectively, close to the estimates based on Ramey et al.'s (2019) data (i.e., for recognition confidence, $r = 0.87$; for SPR, $r = 0.59$).

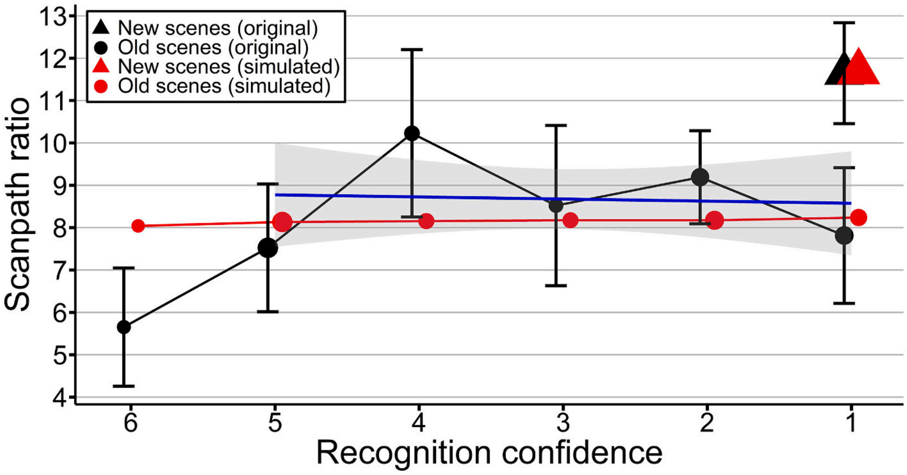


Fig. A.2. Mean scanpath ratio (*SPR*) across different levels of recognition confidence during the test phase in Ramey et al.'s (2019) experiment, and simulation results when between-participants variance is incorporated. The blue regression line is the prediction of *SPR* across different levels of recognition confidence (excluding ratings of 6) for old scenes using the original data. Symbol sizes represent the relative proportions of scenes across different levels of recognition confidence. The shaded area represents the 95% confidence band of the regression line, and the error bars represent the 95% confidence intervals of the *SPR* means. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A.3. Conscious and unconscious effects on SPR and FSA at participant level

Table A.1 presents the difference between conscious and unconscious effects on SPR and FSA respectively at participant level. The difference scores are computed based on subtracting the unconscious effect from the conscious effect for both eyetracking measures. Insomuch as Ramey et al. (2019) concluded that conscious memory uniquely affected FSA while unconscious memory uniquely affected SPR, a negative difference score should be observed for SPR and a positive difference score should be observed for FSA across participants. Only 7 participants exhibited a pattern in line with Ramey et al.'s (2019) conclusion.

Table A.1
Difference scores between conscious and unconscious effects on SPR and FSA at participant level.

Participant	SPR	FSA	Is this pattern in line with the hypothesized dissociation?
1	−1.15	19.82	Yes
2	−0.33	23.62	Yes
3	−6.58	46.07	Yes
4	−10.79	26.10	Yes
5	−0.70	60.10	Yes
6	−1.44	31.76	Yes
7	−4.56	22.08	Yes
8	−3.42	−24.20	No
9	−3.92	−4.77	No
10	−3.51	−54.92	No
11	3.73	104.95	No
12	−1.64	−11.33	No
13	−2.80	−15.34	No
14	2.47	44.01	No
15	−1.56	−29.81	No

A.4. Reliability assessments of critical measures in Experiments 1 and 2

Reliability assessments, based on the random-split approach as discussed in Section 5, were computed for the critical measures in the familiarity judgment stage and the final test stage. Split-half reliability estimates, both uncorrected and Spearman-Brown corrected, are presented in Table A.2. With the exception of the accuracy measure of the 2AFC stage, all measures were of very satisfactory reliability. The lower reliability of the 2AFC measure could be attributed to not only the binary nature of the response (Vadillo et al., 2022), but also the smaller number of trials compared to the recognition rating measures in other tasks.

Table A.2
Split-half reliability estimates based on 5000 random splits across different measures.

Measures	Split-half correlation	Spearman-Brown
Response time (familiarity judgment stage)	0.861	0.925
Recognition confidence rating (familiarity judgment stage)	0.793	0.884
Accuracy (2AFC stage)	0.237	0.378
Recognition confidence rating (final recognition stage)	0.893	0.943

A.5. Alternative analytic approach

The results reported in sections 7.4.2–7.4.3 are largely in line with our *a priori* predictions. There are nevertheless some potential concerns that warrant scrutiny. First, different statistical approaches were adopted, such that while crossed random effects of participant and image were adjusted for when assessing the association between RT and familiarity, these random effects were not adjusted for in other critical comparisons involving paired sample *t*-tests. Second, in tests where RT was treated as a dependent variable, the inherent skewness of its distribution could potentially exert a non-negligible influence on the model estimates, even when a LMM was conducted (Lo & Andrews, 2015; see also Schielzeth et al., 2020).

To address these potential concerns, we replicated all the planned statistical tests using an alternative approach. First, all RT data were subject to a natural log transformation from the outset. Afterwards, outlying log-RTs were identified and removed with respect to each of the image types (i.e., old and new images) based on the standard boxplot method. Specifically, log-RTs over 1.5 interquartile ranges (IQR) from the third quartile and those below 1.5 IQRs from the first quartile were deemed as outliers. As a result, 383 trials, amounting to 2.32% of total trials in the familiarity judgment stage, were excluded. After removal of outliers, log-RT distributions of both image types were largely symmetrical (skewness for old images: 0.41; skewness for new images: 0.44). Then, we conducted all the planned statistical analyses again, but this time using LMMs while adjusting for crossed random effects of participant and image across the board. Crucially, where applicable, log-RT was treated as the dependent variable in place of RT.

Regarding the familiarity judgment stage, the overall contextual cueing effect remained significant, $t(15830) = 6.86$, $p < .001$, $d_z = 0.05$, and the cueing effect restricted to images rated 1 was also significant, $t(3501) = 2.13$, $p = .02$, $d_z = 0.04$. No significant association between log-RT and familiarity (rating 1–5) was found, $b = 0.004$, $t(9057) = 1.19$, $p = .12$. As regards the 2AFC stage of Experiment 1, higher-than-chance accuracy was revealed across participants as before, $t(1406) = 3.57$, $p < .001$, $d_z = 0.10$. It is worth noting that as both old and new images were presented in any given trial during this stage, in addition to the random effect of participant, the random effects of both old and new image were adjusted for. Notwithstanding this, the LMM result indicated a significant fixed effect. Finally, with respect to the final recognition stage of Experiment 2, there was again a significant difference between old and new images that received ratings of 1 in the previous stage, $t(1719) = 3.93$, $p < .001$, $d_z = 0.09$. All in all, the results based on the alternative analytic approach echoed the results of the planned statistical tests reported in the previous sections, thus evidencing their validity.

A.6. Contextual cueing effect at the item level

To further characterize the cueing effect in our study, we examined the effect at the item level, namely the respective cueing effects borne by individual images. Specifically, for each image, its search RT when used as a new image was compared to its search RT as an old image during the familiarity judgment stage. As illustrated in Fig. A.3, when all trials were included, a cueing effect was found in the small majority of images. Nevertheless, a substantial number (36 images, or 37.5% of the image set) did not evoke faster responses. This suggests that the cueing effect in this experimental task may not be as robust as the cueing effect evoked in the conventional array-based contextual cueing paradigm (Chun & Jiang, 1998, 2003).

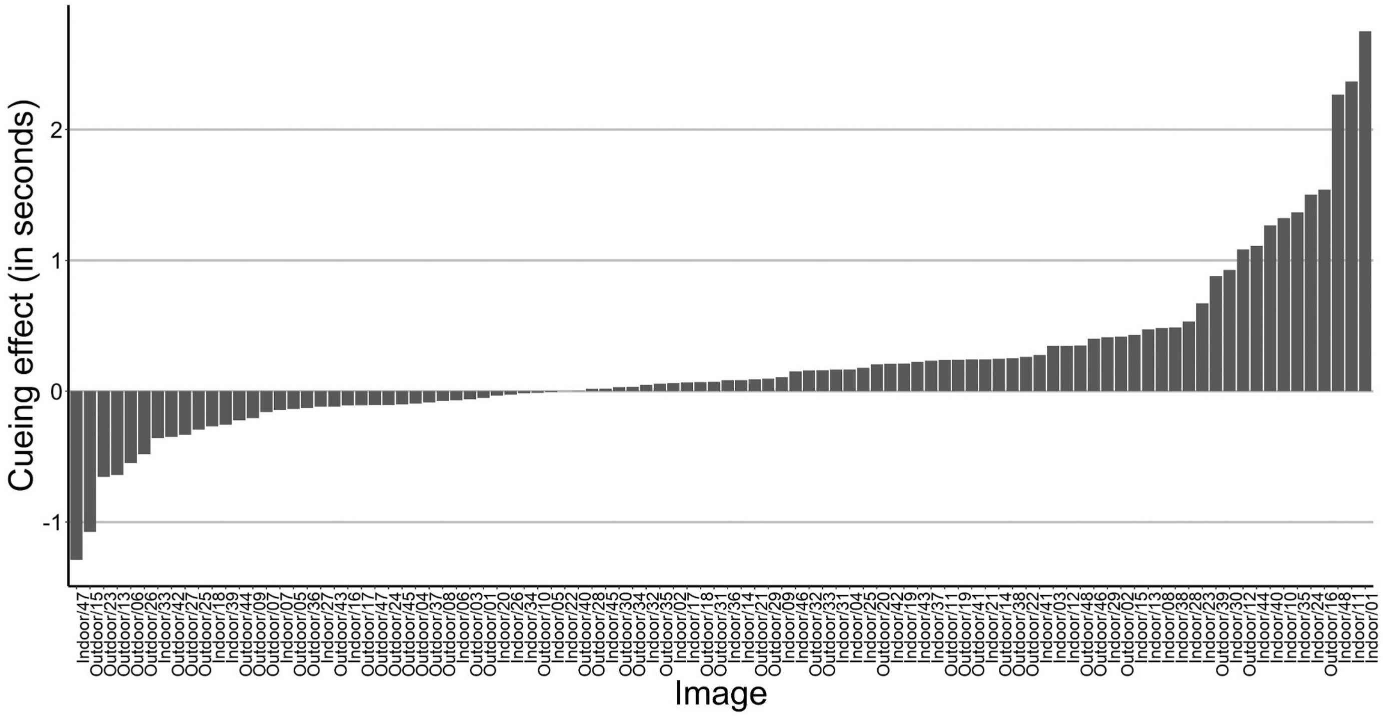


Fig. A.3. Contextual cueing effect in terms of response time (RT) across individual images. Positive values indicate that the search RT for the given image was higher when used as a new image than as an old image during the familiarity judgment stage, corresponding to the standard cueing effect.

A.7. Trial count of image pairs rated 1 across different accuracies

We further investigated whether participants' accuracy in correctly identifying old images across image pairs varied systematically with the number of trials encountered in the 2AFC task. As illustrated in Fig. A.4, for image pairs rated 1, there was a largely even distribution of trial count across different levels of accuracy, except for extreme accuracies (i.e., zero accuracy and perfect accuracy), where the participants concerned encountered relatively few trials. Overall, no systematic variation between number of trials and accuracy was observed.

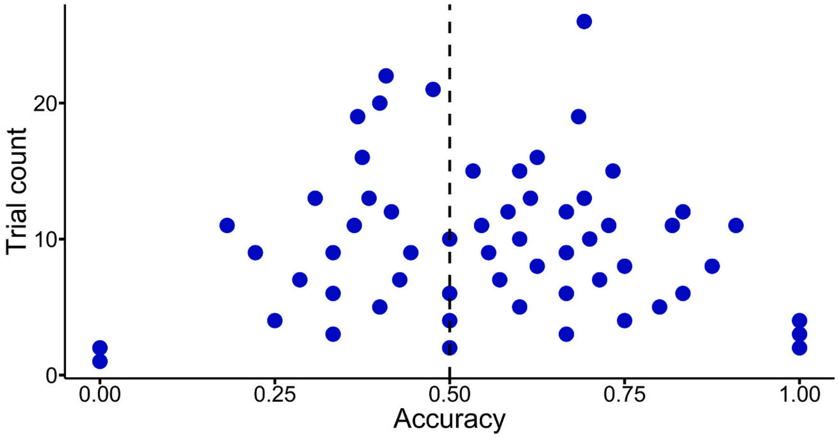


Fig. A.4. Mean accuracy for different trials counts, quantified as the number of pairs rated 1 that were presented in the 2AFC stage. Each point represents the given participant's mean accuracy with the corresponding trial count.

References

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*(456). https://doi.org/10.3389/fpsyg.2017.00456

Bastin, C., & Van der Linden, M. V. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology, 17*, 14–24. https://doi.org/10.1037/0894-4105.17.1.14

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

Bayley, P. J., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2008). Yes/no recognition, forced-choice recognition, and the human hippocampus. *Journal of Cognitive Neuroscience, 20*, 505–512. https://doi.org/10.1162/jocn.2008.20038

Berry, C. J., Kessels, R. P. C., Wester, A. J., & Shanks, D. R. (2014). A single-system model predicts recognition memory and repetition priming in amnesia. *Journal of Neuroscience, 34*(33), 10963–10974. https://doi.org/10.1523/JNEUROSCI.0764-14.2014

Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A unitary signal-detection model of implicit and explicit memory. *Trends in Cognitive Sciences, 12*, 367–373. https://doi.org/10.1016/j.tics.2008.06.005

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. A. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review, 119*, 40–79. https://doi.org/10.1037/a0025464

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40*, 227–259. https://doi.org/10.1006/cogp.1999.0731

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*, 28–71. https://doi.org/10.1006/cogp.1998.0681

Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 29*, 224–234. https://doi.org/10.1037/0278-7393.29.2.224

Colagiuri, B., & Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin and Review, 23*, 1996–2009. https://doi.org/10.3758/s13423-016-1063-0

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–574. https://doi.org/10.1177/00131640121971374

Dunn, J. C. (2003). The elusive dissociation. *Cortex, 39*, 177–179. https://doi.org/10.1016/S0010-9452(08)70096-0

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*, 175–191. https://doi.org/10.3758/BF03193146

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. California, USA: Peninsula Publishing.

Hamann, S. B., & Squire, L. R. (1997). Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience, 111*, 850–854. https://doi.org/10.1037/0735-7044.111.4.850

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hannula, D. E., Althoff, R. R., Warren, D. E., Riggs, L., Cohen, N. J., & Ryan, J. D. (2010). Worth a glance: Using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in Human Neuroscience, 4*. https://doi.org/10.3389/fnhum.2010.00166

Hannula, D. E., Baym, C. L., Warren, D. E., & Cohen, N. J. (2012). The eyes know: Eye movements as a veridical index of memory. *Psychological Science, 23*, 278–287. https://doi.org/10.1177/0956797611429799

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340. https://doi.org/10.1037/0096-3445.110.3.306

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010). Global similarity predicts dissociation of classification and recognition: Evidence questioning the implicit-explicit learning distinction in amnesia. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*, 1529–1535. https://doi.org/10.1037/a0020598

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Jiang, Y. V., Sisk, C. A., & Toh, Y. N. (2019). Implicit guidance of attention in contextual cueing: Neuropsychological and developmental evidence. *Neuroscience and Biobehavioral Reviews, 105*, 115–125. https://doi.org/10.1016/j.neubiorev.2019.07.002

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced- choice recognition tests in healthy subjects and amnesics. *Neuropsychologia, 38*, 1333–1341. https://doi.org/10.1016/S0028-3932(00)00055-5

Kroell, L. M., Schlagbauer, B., Zinchenko, A., Müller, H. J., & Geyer, T. (2019). Behavioural evidence for a single memory system in contextual cueing. *Visual Cognition, 27*, 551–562. https://doi.org/10.1080/13506285.2019.1648347

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26. https://doi.org/10.18637/JSS.V082.I13

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*(863). https://doi.org/10.3389/fpsyg.2013.00863

Lange, N., Berry, C. J., & Hollins, T. J. (2019). Linking repetition priming, recognition, and source memory: A single-system signal-detection account. *Journal of Memory and Language, 109*. https://doi.org/10.1016/j.jml.2019.104039

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01171

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 3–26. https://doi.org/10.1037//0097-7403.28.1.3

Malejka, S., Vadillo, M. A., Dienes, Z., & Shanks, D. R. (2021). Correlation analysis to investigate unconscious mental processes: A critical appraisal and mini-tutorial. *Cognition, 212*. https://doi.org/10.1016/j.cognition.2021.104667

Mee, R. W., & Chua, T. C. (1991). Regression toward the mean and the paired sample t test. *The American Statistician, 45*, 39–42. https://doi.org/10.1080/00031305.1991.10475763

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process: Implications for dual-process theories of recognition memory. *Psychological Science, 20*, 509–515. https://doi.org/10.1111/j.1467-9280.2009.02324.x

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313. https://doi.org/10.1093/comjnl/7.4.308

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences, 12*, 285–290. https://doi.org/10.1016/j.tics.2008.04.009

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B. H. Ross (Ed.), *Vol. 54. The psychology of learning and motivation* (pp. 167–215). Academic Press.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*, 1105–1107. https://doi.org/10.1038/nn.2886

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science, 9*, 247–255. https://doi.org/10.1111/1467-9280.00051

Palfi, B., & Dienes, Z. (2020). Why Bayesian "evidence for H1" in one condition and Bayesian "evidence for H0" in another condition does not mean good-enough Bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science, 3*, 300–308. https://doi.org/10.1177/2515245920913019

Parsons, S. (2020). *splithalf; robust estimates of split half reliability [Computer software]*. Retrieved from https://doi.org/10.6084/m9.figshare.5559175.v5.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Poldrack, R. A. (1996). On testing for stochastic dissociations. *Psychonomic Bulletin and Review, 3*, 434–448. https://doi.org/10.3758/BF03214547

R Core Team. (2022). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org.

Ramey, M. M., Yonelinas, A. P., & Henderson, J. M. (2019). Conscious and unconscious memory differentially impact attention: Eye movements, visual search, and recognition processes. *Cognition, 185*, 71–82. https://doi.org/10.1016/j.cognition.2019.01.007

Richardson-Klavehn, A., Clarke, A. J. B., & Gardiner, J. M. (1999). Conjoint dissociations reveal involuntary "perceptual" priming from generating at study. *Consciousness and Cognition, 8*, 271–284. https://doi.org/10.1006/ccog.1999.0382

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Thousand Oaks, CA: SAGE Publications.

Rothkirch, M., Shanks, D. R., & Hesselmann, G. (2022). The pervasive problem of post hoc data selection in studies on unconscious processing: A reply to Sklar, Goldstein, and Hassin (2021). *Experimental Psychology, 69*, 1–11. https://doi.org/10.1027/1618-3169/a000541

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science, 11*, 454–461. https://doi.org/10.1111/1467-9280.00288

Savage, M. (2018). Reliability, split-half. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (p. 1421). Thousand Oaks, CA: SAGE Publications.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501–518. https://doi.org/10.1037/0278-7393.13.3.501

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*, 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science, 16*, 396–414. https://doi.org/10.1177/1745691619863798

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin and Review, 24*, 752–775. https://doi.org/10.3758/s13423-016-1170-y

Shanks, D. R., & Berry, C. J. (2012). Are there multiple memory systems? Tests of models of implicit and explicit memory. *Quarterly Journal of Experimental Psychology, 65*, 1449–1474. https://doi.org/10.1080/17470218.2012.691887

Shanks, D. R., & Perruchet, P. (2002). Dissociation between priming and recognition in the expression of sequential knowledge. *Psychonomic Bulletin and Review, 9*, 362–367. https://doi.org/10.3758/BF03196294

Smith, C. N., & Squire, L. R. (2017). When eye movements express memory for old and new scenes in the absence of awareness and independent of hippocampus. *Learning and Memory, 24*, 95–103. https://doi.org/10.1101/lm.043851.116

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Spering, M., & Carrasco, M. (2015). Acting without seeing: Eye movements reveal visual processing without awareness. *Trends in Neurosciences, 38*, 247–258. https://doi.org/10.1016/j.tins.2015.02.002

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*, 232–243. https://doi.org/10.1162/jocn.1992.4.3.232

Squire, L. R., & Dede, A. J. (2015). Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology, 7*(3). https://doi.org/10.1101/cshperspect.a021667

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301–306. https://doi.org/10.1126/science.2296719

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin and Review, 23*, 87–102. https://doi.org/10.3758/s13423-015-0892-6

Vadillo, M. A., Malejka, S., Lee, D. Y. H., Dienes, Z., & Shanks, D. R. (2022). Raising awareness about measurement error in research on unconscious mental processes. *Psychonomic Bulletin & Review, 29*, 21–43. https://doi.org/10.3758/s13423-021-01923-y

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517. https://doi.org/10.1006/jmla.2002.2864

Zaki, S. R., & Nosofsky, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective, & Behavioral Neuroscience, 1*, 344–359. https://doi.org/10.3758/CABN.1.4.344