Article:
Berens, Sam, Richards, Blake and Horner, Aidan James orcid.org/0000-0003-0882-9756 (2020) Dissociating memory accessibility and precision in forgetting. Nature Human Behaviour. 866–877. ISSN 2397-3374

https://doi.org/10.1038/s41562-020-0888-8

Dissociating memory accessibility and precision in forgetting

Sam C Berens[1,2,*], Blake A Richards [3,4,5], & Aidan J Horner[1, 7,*]

1. Department of Psychology, University of York, UK.
2. School of Psychology, University of Sussex, UK.
3. Mila, Montréal, QC, Canada.
4. Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada.
5. School of Computer Science, McGill University, Montréal, QC, Canada.
6. Learning in Machines and Brains Program, Canadian Institute for Advanced Research, Toronto, ON, Canada.
7. York Biomedical Research Institute, University of York, UK.

* Corresponding authors: s.berens@sussex.ac.uk & aidan.horner@york.ac.uk

Author ORCID identifies:

Sam Berens: https://orcid.org/0000-0001-8197-8745
Blake Richards: https://orcid.org/0000-0001-9662-2151
Aidan Horner: https://orcid.org/0000-0003-0882-9756

Abstract

Forgetting involves the loss of information over time, however, we know little about what form this information loss takes. Do memories become less precise over time, or do they instead become less accessible? We assessed memory for word-location associations across 4 days, testing whether forgetting involves losses in precision vs accessibility and whether such losses are modulated by learning a generalisable pattern. We show that forgetting involves losses in memory accessibility with no changes in memory precision. When participants learnt a set of related word-location associations that conformed to a general pattern, we saw a strong trade-off; accessibility was enhanced whereas precision was reduced. However, this trade-off did not appear to be modulated by time or confer a long-term increase in the total amount of information maintained in memory. Our results place theoretical constraints on how models of forgetting and generalisation account for time-dependent memory processes.

Introduction

Forgetting is an inevitable consequence of remembering. We forget many of our everyday experiences over time, remembering only a small proportion of the large volume of information we process on a daily basis[1]. Psychologists have focussed on characterising the rate at which forgetting occurs – epitomised by Ebbinghaus' forgetting curves[2], and have asked why it occurs – for example via interference or decay[3–6]. This focus on both the when and why of forgetting has perhaps been at the expense of asking what is forgotten. When forgetting occurs, what type of information is lost? This question is critical given the proposal that forgetting is beneficial to decision-making processes[7,8]. If we are to understand how forgetting aids decision-making, we first need to reveal the form that such forgetting takes.

Here we outline two possible ways in which forgetting might occur – via decreases in memory accessibility or precision. Imagine being in a park and meeting a friend by a fountain in the north-east corner. Sometime in the future, you might want to remember the specific location where you met. A decrease in accessibility would mean a reduced probability of retrieving that specific memory. However, if successfully retrieved, you may remember the meeting location with the same accuracy as before. A decrease in precision would mean that the probability of successful retrieval does not change, but the spatial precision of retrieval does decrease. You might remember meeting your friend in the park, but not specifically by the fountain in the north-east corner. Both accessibility and precision can be defined as a loss of information, yet these two types of information loss should be behaviourally dissociable. Further, these two potential forms of forgetting might be underpinned by distinct mechanisms. For example, whereas accessibility might change as a function of the connection strength between a retrieval cue and its associated memory trace, precision might change as a function of noise in the underlying trace itself. Note, here we define 'forgetting' broadly in terms of a loss of information, as opposed to a more restrictive definition in relation to whether retrieval has been successful or not.

A number of theoretical accounts suggest that forgetting should involve different rates of decline for certain types of mnemonic information. In particular, Fuzzy-trace theory (FTT) posits that episodic memories are encoded by two independent traces that may be stored and retrieved in parallel[9]. One of these traces represents the fine-grained details of an event whereas the other encodes gist information in the form of semantic features. Relatedly, building on multiple trace theory[10], the Trace Transformation Theory (TTT) proposes that the hippocampus supports the encoding and retention of episodic, context-rich, memories, while the neocortex transforms such representations into more semantic, gist-like, memories[11,12]. Empirical observations support these dissociations by showing that

perceptual details may be lost faster than gist information[13–16]. However, this research focusses on loss of information for two distinct mnemonic representations, as opposed to losses in accessibility and precision for individual memory representations.

Recent research has shown that accessibility and precision are perhaps distinct components of an episodic representation. First, although accessibility and precision positively correlate across participants, they each have unique variance[17]. Participants can make accurate metacognitive judgements at retrieval related to this unique variance – they can subjectively report how accessible and precise memory retrieval is on a trial-by-trial basis[18]. Accessibility and precision have also been shown to be neurally dissociable. fMRI evidence has shown that trial-by-trial accessibility correlates with hippocampal activity, whereas trial-by-trial precision correlates with angular gyrus activity[19] (but also see[20]). Further, repetitive transcranial magnetic stimulation to the lateral parietal cortex produces improvements in precision, but not accessibility[21]. Thus, although there is evidence from working memory paradigms that accessibility and precision can be characterised using a single parameter model[22], long-term memory studies have provided evidence that accessibility and precision are (at least partially) behaviourally and neurally dissociable.

One previous study has specifically focused on accessibility and precision in relation to forgetting in working memory[23]. Sun *et al* showed that encoding similar interfering material led to decreases in precision (referred to as 'blurring'), whereas less similar material led to decreases in accessibility (referred to as 'erasure'). In contrast to Sun *et al*, who focus on experimental interference in working memory, we focus on whether these plausibly distinct long-term memory processes can be dissociated via their forgetting rates over time. Assessing the temporal profile of forgetting is critical given that this reflects more naturalistic 'everyday' forgetting (i.e., participants are free to go about the daily lives in between encoding and retrieval). If forgetting does play a role in optimising decision-making processes, knowing what information is available to these processes, and when it is available, is critical to the development of models of memory-guided decision-making. Additionally, understanding whether forgetting principally involves losses in precision or accessibility will inform theoretical accounts of long-term memory retention.

To date, research into forgetting has predominantly used binary measures of memory retrieval, where each retrieval trial can be classified as either correct or incorrect[5]. Forgetting under these experimental conditions is typically assessed by comparing accuracy (i.e., the proportion of correct responses) across experimental conditions. This general approach has been highly successful in delineating interference versus decay accounts of forgetting, and recently has shown that item-based

familiarity is more susceptible to interference than decay, whereas recollection is more susceptible to decay than interference[6]. However, this experimental approach is not capable of dissociating between accessibility and precision. Note, there is no clear correspondence between familiarity vs recollection, and accessibility vs precision. Indeed, accessibility and precision may be independent components of recollection (dependent on the experimental task used). As such, we make no claims in relation to the debate surrounding possible dissociations between familiarity and recollection, instead focussing on potential dissociations between accessibility and precision.

As noted, 'precision' measures of memory have been used to study both working memory[24,25] and long-term memory[18–20]. Here participants are required to remember a continuous perceptual detail of a stimulus, such as the colour of an object or its location on a circle. In the long-term memory literature, it is typical to pair a word with a location on a circle at encoding such that participants learn a 'word-location association'[17,18,26]. At retrieval, the word acts as the cue and participants have to move a cursor to the remembered location on the circle. Memory 'precision' is measured as the angular difference (error) between the correct and remembered location. Thus, memory performance is assessed with a continuous rather than binary measure.

Precision memory measures have also been combined with a statistical approach (mixture modelling) that allows for the characterisation of both memory accessibility and precision. Taking the angular error across all trials, mixture models allow one to fit a circular bell-shaped distribution (a von Mises distribution) to the data. Once fit, the width of the von Mises distribution reflects the precision of memory retrieval. For example, if a participant is remembering circular locations very precisely, the distribution of angular errors will be narrow. Memory accessibility can also be estimated by considering the proportion of angular errors that were likely generated by the von Mises distribution, rather than being uniformly distributed around the circle (indicative of guessing). Importantly, these measures of accessibility and precision are independent of each other, such that if precision is high, accessibility can be either high or low (and vice versa). The combination of precision memory measures and mixture modelling therefore offers a unique opportunity to assess the extent to which forgetting decreases accessibility or precision.

Current measures of accessibility and precision are, to date, not directly comparable. Whereas the accessibility measure is related to 'proportion correct' in a more typical memory experiment, the precision measure relates to the width of the fitted von Mises distribution. To assess the extent to which forgetting is characterised by decreases in accessibility or precision, we need to develop a common metric. The concept of 'information loss' is related to entropy, which measures the lack of

predictability in a given system[27]. As information is lost, the system behaves less predictably, and so responses will become more variable. Here we use the entropy of behavioural responses[28,29] to measure the amount of information loss across time. We introduce a common metric to directly compare information loss in terms of both accessibility and precision. Using this common metric, we measured the accessibility and precision of word-location associations in an online experiment involving a large sample of participants. Specifically, we tracked changes in accessibility and precision across time by allocating participants to one of 7 retention interval conditions such that retrieval occurred either 0 hrs, 3 hrs, 6 hrs, 12 hrs, 24 hrs, 48 hrs, or 96 hrs after initial encoding. We directly compared the pattern of decreases in accessibility and precision across these intervals. Thus, we were able to assess whether accessibility and precision decreased at differing rates.

Episodic memories are not encoded in isolation. We often experience events that are highly related, and can use that overlapping content to generalise across a set of events (referred to as schema[30–32]). Theories of consolidation, such as Standard Consolidation Theory[33] (SCT) and TTT (introduced above[11,12]) propose that schematic representations, supported by the neocortex, are more stable and resilient to forgetting relative to more specific, hippocampal-based, episodic representations. Although existing schema can support the encoding of new item-based information[34], the ability to generalise across related experiences might come at the expense of remembering individual events precisely[35]. Recent evidence suggests that participants use schema when making mnemonic decisions (which may be further modulated by systems consolidation[36]), and that this can result in systematic biases towards the 'average' representation across events when recalling individual events[37]. Thus, generalisation across a set of related experiences may result in a trade-off – decreasing total information loss over time at the expense of losing precise information related to specific events.

Here we asked whether similar events alter the rate of information loss for accessibility and precision over time. Word stimuli in the experiment were grouped into two semantic categories, 'manmade' and 'natural'. Participants then associated these words with different locations around a circle (Figure 1A). The circular locations for one group of words were entirely random at encoding. Locations associated with the other group of words were spatially clustered (according to an underlying von Mises distribution with a fixed-width; conceptually similar to Richards *et al*[38]). At test, participants were asked to reproduce the location associated with each word (Figure 1B). This clustering of locations for semantically similar words may allow participants to generalise across a set of related experiences (either at encoding or retrieval), potentially altering the rate of information loss for accessibility or precision (see hypotheses below and pilot data in the Supplementary Information). The present study aimed to systematically characterise differential losses of accessibility and precision

over the 7 retention intervals. However, future work is needed to reveal whether such losses are driven by processes at encoding or retrieval, and what the nature of the underlying representations are in the clustered and non-clustered condition.

Using online testing, we tracked rates of information loss in terms of accessibility and precision for word-location associations that were either randomly distributed around a circle (non-clustered) or spatially clustered. Our experimental approach focused on memory for the word-location associations, rather than item memory for individual words (see planned exploratory analyses that differentiate item and associative memory).

Our preregistered analyses tested five specific hypotheses (each was assessed in our pilot data, providing evidence in favour of the alternative hypothesis; *BFs* > 6; see Supplementary Figure 1 and Supplementary Table 1). Before examining separate measures of accessibility and precision, we made two predictions in relation to the total amount of information (I_t, see Methods). I_t is a measure of the total amount of information in a given condition that takes into account the level of both accessibility and precision. First, we predicted a decrease in total information across time, specifically for non-clustered words, consistent with the presence of forgetting (Hypothesis 1). Second, we predicted that clustered words would confer an overall memory benefit relative to non-clustered words, consistent with a benefit when schema are formed (regardless of time; Hypothesis 2). These hypotheses act as positive controls, providing greater certainty for the validity of the more specific hypotheses below.

Of central theoretical interest was whether accessibility and precision differ in relation to forgetting, and how this further interacts with our manipulation of clustering. Here we decomposed the measure of total information (I_t) into separate measures of accessibility (I_p) and precision (I_k; the subscripts p and k refer to the corresponding parameters in the mixture model). First, we predicted that the temporal profile of forgetting, specifically for non-clustered words, differs for accessibility and precision as these measures reflect different components of memory (Hypothesis 3). We remained agnostic as to whether this forgetting rate will be faster or slower for accessibility vs precision.

Our final two preregistered hypotheses related to how clustering differentially affects accessibility and precision. As previously discussed, computational work has suggested a trade-off between generalisation and remembering individual events precisely[35]. Theories of consolidation also predict that gist-like, schematic, representations should be retained for longer periods of time, and that these representations might aid memory accessibility that the expense of precision[11,12]. We therefore predicted that accessibility and precision will differ between the clustered and non-clustered condition (regardless of time; Hypothesis 4). In particular, this interaction was likely to present as

increased accessibility, but decreased precision, in the clustered relative to the non-clustered condition (see pilot data). However, the statistical test for this was chosen to be non-directional. Furthermore, this interaction was predicted to be modulated by time, such that the rate of information loss for accessibility vs precision would differ dependent on whether words were clustered or non-clustered (Hypothesis 5). This three-way interaction was predicted to present as a more rapid loss in accessibility in the non-clustered (relative to clustered) condition, in contrast to a more rapid loss in precision in the clustered (relative to non-clustered) condition (see pilot data). Again, however, the statistical test for this was non-directional.

As mentioned above, our principal hypotheses and preregistered analyses do not differentiate between failures to recognise individual cue words, and failures to recall specific locations when a cue word is remembered. Nonetheless, potentially dissociating between these possibilities is also important. As such, at the end of each word-location retrieval trial, participants were asked to provide subjective judgments regarding whether they remembered both the cued word and its associated location (associative retrieval), the word alone (item recognition), or neither (see Figure 1B).

Planned exploratory analyses then tested for possible dissociations between item- vs associative-memory. These analyses provided the potential to shed light on differences between the clustered and non-clustered conditions. For instance, a performance advantage for clustered trials could result from either: (1) better memory for specific word-location associations within a spatial schema (enhanced retention), or (2) mnemonic generalisation involving the retrieval of representative locations when specific word-location associations have been forgotten (i.e., exemplar or prototype generalisation[39]). Higher proportions of associative retrieval judgments in the clustered condition would support an enhanced retention account whereas lower proportions would suggest the use of generalisation. Thus, our post-trial question provided some measure of whether specific words or word-location associations are forgotten, depending on whether they are part of a semantic cluster.

To summarise, we used online testing, precision memory measures, and mixture modelling to assess forgetting across time. Using a common metric (information), we directly compared decreases in accessibility and precision over time and investigated how these decreases were modulated by overlapping experience (i.e., clustered vs non-clustered words).

Results

Final sample size and demographics

In accordance with our recruitment protocol, we collected data until the Bayes factors for each of our preregistered hypotheses became sensitive (indicating 10 times more evidence in favour of either the null or alternative hypothesis), or our maximum feasible sample size had been obtained. In fact, data collection stopped at the maximum sample size with all but one of the Bayes factors (Hypothesis 5) reaching our sensitivity threshold. In total, the final sample included data from 431 participants with 60, 68, 62, 60, 61, 60, and 60 participants in the 0, 3, 6, 12, 24, 48, and 96 hrs conditions respectively. The variation in final sample sizes across the retention intervals was driven by logistic difficulties in knowing whether recruited participants were likely to provide complete datasets, and thus over-recruiting in some retention intervals. Participants' ages were uniformly distributed between our upper and lower age limits (18-35 inclusive; median age: 28 years) and approximately 65% identified as female.

Preregistered analyses

Effect sizes, Bayes factors, and frequentist statists for each of our pre-registered hypotheses are presented in Table 1. Figure 2A displays each measure of mnemonic information (I_t, I_p and I_k) with individual data points and mean estimates from generalised-linear mixed-effects models of the data (GLMMs; see Methods). Additionally, Figure 2B plots kernel density estimates (averaged across participants) that characterise the distribution of angular errors in each condition. These are produced using a non-parametric technique and so provide an alternative means of visualising changes in performance independent of the mixture models that were used to compute I_t, I_p and I_k (see Supplementary Information for details). Raw means and standard deviations of each outcome measure are provided in Supplementary Table 2.

As predicted, the total amount of information retained in memory (I_t) decreased across retention intervals in the non-clustered condition, consistent with the forgetting of word-location associations (BF_{10} = 117***;*** Hypothesis 1). However, counter to our predictions, we did not find an overall difference in performance between the clustered and non-clustered conditions (i.e. a main effect; Hypothesis 2). Indeed, the Bayes factor for this test shows substantial evidence in favour of the null hypothesis (BF_{10} = 0.054). Thus, we provide evidence of forgetting, as measured by decreases in total information in the non-clustered condition over time, and evidence for no overall memory benefit for the clustered relative to non-clustered condition (as measured by total information, regardless of time).

Hypothesis 3 related to our prediction that accessibility (I_p) and precision (I_k) decline at different rates and this is strongly supported by our analyses (BF_{10} = 35.1). Specifically, estimates of memory accessibility declined much more rapidly than estimates of memory precision. Hypothesis 4 concerned our prediction that clustering would differentially alter the levels of accessibility and precision in general (i.e. when averaging across retention intervals). Again, this was strongly supported by our analyses with clustered trials eliciting higher levels of accessibility, but lower levels of precision (BF_{10} = 5.98×10^6; c.f. I_k estimates in the 0 hrs and 3 hrs conditions). We also predicted that this differential effect of clustering would be further modulated by time (Hypothesis 5). In particular, we expected to see a more rapid loss of accessibility in the non-clustered condition, and a more rapid loss of precision in the clustered condition. Despite this, the Bayes factor for Hypothesis 5 revealed there to be over 5 times more evidence in favour of no effect (BF_{10} = 0.188). While this result does not reach our sensitivity threshold, it implies that the rates of change in accessibility and precision are not substantially altered by the clustering manipulation.

Exploratory analyses

Post-hoc tests

The two key positive findings from our pre-registered hypotheses were: *(1)* accessibility and precision decline at different rates in the non-clustered condition (Hypothesis 3) and *(2)* accessibility and precision are modulated by the clustering manipulation, regardless of time interval (Hypothesis 4). Exploratory analyses to characterise these interactions were conducted. In relation to Hypothesis 3, we tested for evidence of exponential losses in accessibility and precision separately. This revealed strong evidence for a decline in accessibility in the non-clustered condition (d = 0.279, BF_{10} = 5620, t_{1696} = 4.762, p < .001), but evidence in favour of the null (i.e., no decline) for precision (d = 0.002, BF_{10} = 0.0612, t_{1696} = 0.057, p = .955). We therefore provide clear evidence that forgetting in this experimental paradigm is driven solely by loses in accessibility and not precision (at least in the non-clustered condition). Finally, in relation to Hypothesis 4, we found strong evidence for increased accessibility (d = 0.117, BF_{10} = 4.75 × 10^5, t_{1696} = 5.771, p < .001) but decreased precision (d = 0.061, BF_{10} = 225, t_{1696} = 4.317, p < .001) in the clustered relative to the non-clustered condition (collapsed across time interval). The clustering manipulation therefore increased accessibility at the expense of precision. As such, we provide strong evidence for two independent effects: *(1)* decreased accessibility but not precision across time intervals (in the non-clustered condition) and *(2)* increased accessibility and decreased precision in the clustered relative to the non-clustered condition, regardless of time interval.

The pre-registered analyses provided strong evidence against Hypothesis 2; that clustered trials should generally yield higher levels of performance as measured by I_t. Nonetheless, clustered I_t scores were notably larger than non-clustered I_t scores in the 0 hrs and 3 hrs conditions (*d* = 0.106 and 0.164 respectively). Additionally, while our pre-registered analyses found no main effect of clustering, our pilot data strongly suggested that one should be found (see Supplementary Information). Given this, we wished to explore whether the main effect we originally observed in the pilot was best characterised as an interaction between clustering and delay. We therefore tested for this interaction with the same linear contrast used in the main analyses. This produced weak evidence that clustering yields higher levels of performance at shorter retention intervals, with a smaller or non-existent effect at longer delays; d = 0.312, BF_{10} = 5.69, t_{848} = 2.771, p = .006. Thus, the clustered condition may confer memory benefits over shorter time intervals, with this advantage possibly deceasing across time. This will need to be tested in planned confirmatory analyses.

Planned exploratory analyses

We first explored whether there were systematic changes in the subjective memory judgments that participants provided after each test trial. Here, participants indicated whether they remembered the word-location association, the word alone, or neither. As planned, we specified a cumulative link mixed-effects regression model to predict changes in the proportion of test trials that received either a 'Word + location', 'Word only', or 'Neither' response. Figure 3 plots the model-derived probability estimates for each response type across conditions. The model indicated strong evidence for time-dependent decreases in the subjective retrieval of words and word-location associations in both the non-clustered and clustered conditions; BF_{10} = 2484, z = 4.919, p < .001, and BF_{10} = 574, z = 4.468, p < .001 (respectively; tested by the same linear contrast used in our pre-registered hypotheses). This manifested as a reduced proportion of 'Word + location' responses at longer retention intervals, marginally fewer 'Word only' responses at the same time-points, and corresponding increases in 'Neither' responses. We also found evidence for a main effect of clustering indicating that, on average, clustered trials received more 'Word + location' and 'Word only' responses (irrespective of retention interval); BF_{10} = 6.97, z = 2.900, p = .004. This effect was principally driven by differences in 4 of the delayed retention intervals (specifically, 3 hrs, 12 hrs, 48 hrs, and 96 hrs). However, there was no evidence for a consistent interaction between clustering and retention interval; BF_{10} = 0.592, z = 1.629, p = .103 (nor evidence in favour of the null hypothesis of no interaction).

We also re-ran the main analyses testing each pre-registered hypothesis after excluding the set of all test trials that received a 'Neither' response (i.e. removing trails where the cue word was not subjectively recognised). This analysis aimed to test whether losses in memory accessibility reflect

either: *(1)* reduced accessibility for the cue word *per se*, or *(2)* failures to maintain the word-location association (in the presence of item memory for the word). As this analysis was performed on a restricted subset of trials, the mixture models that provided estimates of performance could not be adequately fitted to the data for all participants in the main sample. As such, this analysis included data from only 319 participants with 56, 53, 49, 44, 44, 42, and 31 participants in the 0, 3, 6, 12, 24, 48, and 96 hrs conditions respectively.

The results were largely similar to those reported above (effect sizes, Bayes factors, and frequentist statistics detailed in Supplementary Table 3). Importantly, even after excluding 'Neither' trials, we still observed differential rates of loss for accessibility and precision in the non-clustered condition (Hypothesis 3; d = 0.322, BF_{10} = 7.699, t_{1248} = 2.886, p = .004). As before, this was driven by large reductions in accessibility (I_p) across retention intervals and was evident for both the clustered and non-clustered conditions (pooled effect: d = 0.365, BF_{10} = 14675, t_{1248} = 4.938, p < .001, raw effect size: 0.277 *nats)*. Critically, these decreases are comparable, if not larger than, the analogous effect in the main, pre-registered analysis (d = 0.309, BF_{10} = 6.72 × 10^6, t_{1696} = 6.101, p < .001, raw effect size: 0.260 *nats).* Given this, losses in accessibility appear to be principally driven by failures to maintain the word-location association rather than reduced accessibility for the cue word *per se*. In contrast to the pre-registered analysis, the restricted analysis showed more evidence in favour of Hypothesis 5 rather than the null (d = 0.406 BF_{10} = 4.472, t_{1248} = 2.613, p = .009). As originally hypothesised, this effect suggested that there were consistent time-dependent decreases in memory precision for clustered trials, but no such decreases for non-clustered trials (see Supplementary Figure 2).

Additional exploratory analyses

While our preregistered analyses demonstrated that memory performance decreased over time, we explored whether location responses became increasingly influenced by a spatial schema that represented approximate locations in the clustered condition. To do this, we first produced kernel density estimates that quantified the spatial distribution of participant's responses (similar to Richards *et al*[38]). This was done for clustered and non-clustered test trials separately (condition averages plotted in Supplementary Figure 3). Importantly, the kernel density estimates reflected the absolute position of responses relative to centre of the experimentally imposed cluster, not the accuracy of those responses *per se*. Given these estimates, we then computed the Kullback–Leibler divergence (D_{KL}) between participant responses and the spatial pattern characteristic of clustered locations. D_{KL} scores vary between zero and positive infinity with low values indicating a close correspondence between responses and the clustered pattern. Full details of these estimation steps are provided in the Supplementary Methods.

As in the pre-registered analyses, we then generated a GLMM to model changes in D_{KL} scores as a function of retention interval and the clustering manipulation. Mean estimates of D_{KL} for each condition are plotted in Figure 4. This exploratory GLMM highlighted a large main effect of clustering indicating that, across all retention intervals, responses in the clustered condition were more similar to the underlying spatial pattern than responses in the non-clustered condition (d = 0.482, BF_{10} = 4.92 × 10^{59}, t_{848} = 16.805, p < .001). There was no evidence for a main effect of delay (d = 0.158, BF_{10} = 1.158, t_{848} = 2.247, p = .025; tested by the same linear contrast used in our pre-registered hypotheses). Nonetheless, we did detect a strong clustering by delay interaction (d = 0.364, BF_{10} = 21.921 × 10^{59}, t_{848} = 16.805, p < .001). This reflected the fact that, while D_{KL} scores remained stable in the non-clustered condition (d = 0.071, BF_{10} = 0.165, t_{848} = 0.852, p = .394), scores in the clustered condition increased implying a growing dissimilarity between location responses and the underlying spatial pattern (d = 0.275, BF_{10} = 18.9, t_{848} = 3.247, p = .001). Given this, we explored whether the changes in D_{KL} where strongly related to changes in accessibility and/or precision for the clustered items. To do this, we regressed mean estimates of clustered D_{KL} scores for each retention interval against the corresponding means for I_p and I_k. This showed that timepoint-by-timepoint changes in D_{KL} were almost entirely predicted by linear changes in I_p (partial R^2 = .906). In contrast, the relationship between D_{KL} scores and I_k was much weaker (partial R^2 = .286). This implies that the time-dependent changes in pattern divergence are related to the previously reported decreases in memory accessibility and are only minimally influenced by the small changes in precision.

Discussion

A principal aim of this study was to establish whether the forgetting of long-term associative memories entails losses in memory accessibility, memory precision, or both. Here, participants learnt associations between words and distinct spatial locations around a circle. As predicted, memory for these associations declined over time. Importantly, our results clearly demonstrate that this decline in memory performance predominantly involved losses in memory accessibility for specific word-location associations. At the same time, there were negligible changes in the precision of locations that were correctly retrieved. If a word-location association was successfully accessed, it was retrieved with the same level of precision as at immediate test.

This mirrors recent research on contextually rich event memories suggesting that, while the number of remembered details dramatically reduces with time, details that are remembered can be recalled with remarkable accuracy[43]. Additionally, it has been shown that episodic events are forgotten in an all-or-none manner, where accessibility for the key features of a memory trace decrease uniformly[44]. Together, these results suggest that episodic memories that remain accessible continue to be retrieved in a holistic fashion, and that the constituent features of those memories may be accessed with unchanged levels of precision. It is noteworthy that previous research has shown that encoding overlapping content in a working memory task can lead to losses in precision[23]. In the present study, the learning phase involved encoding a large number of word-location associations. As such, any interference from related material may have led to decreases in precision at the point of encoding (or shortly after). Nonetheless, our results clearly indicate that, following encoding of data that does not contain general patterns, there are no further losses in precision, despite clear evidence of forgetting.

A number of neurobiological mechanisms that may contribute to forgetting have now been identified. These include dopamine-induced signalling cascades, within-neuron receptor transport, and hippocampal neurogenesis[8,45,46]. Our results suggest that these mechanisms may act to reduce the accessibility (or availability) of independent memory traces, while not affecting the precision of traces that remain accessible. This hypothesis is consistent with studies of engram cells in the rodent hippocampus. Specifically, 'silent' memory engrams have been observed which are no longer activated by natural retrieval cues but can be artificially expressed to induce retrieval[47,48]. As such, the precision of an engram may be unrelated to the ease with which it is accessed.

The current study also sought to track the maintenance of overlapping (clustered) associations that may be represented by a generalised pattern or rule. This was achieved by clustering locations for one group of semantically related words (the clustered condition) and comparing memory performance in

this condition to a separate group of semantically related words that were associated with entirely random locations around the circle (the non-clustered condition). We predicted that learning overlapping associations would generally aid performance and increase the apparent level of mnemonic information maintained by participants. This prediction did not hold. Although there was relatively weak evidence of an advantage for clustered words at short retention intervals, this rapidly diminished with time (as seen in exploratory analyses). Indeed, our pre-registered hypothesis testing for greater total information in the clustered relative to the non-clustered condition (irrespective of retention interval) provided strong evidence in favour of the null hypothesis. However, while the amount of total mnemonic information was similar between the clustered and non-clustered conditions, the quality of that information was very different. As predicted, relative to the non-clustered condition, words associated with clustered locations could cue retrieval more frequently across all retention intervals, at the expense of reduced precision (a time-independent trade-off between accessibility and precision). Importantly, given the similar levels of total information in the clustered and non-clustered conditions, we can conclude that this trade-off is a genuine trade-off, with no evidence that the increase in accessibility outweighed the decrease in precision.

This result is consistent with suggestions that extracting patterns across a set of memories aids performance when generalising knowledge at the expense of a loss of detail for specific memory representations[35]. The finding also mirrors working memory studies demonstrating that encoding similar visual features leads to decreases in precision, whereas encoding dissimilar features results in decreased accessibility[23]. As noted above, the reduced precision in these working memory experiments is thought to reflect interference between similar items. Accordingly, it is likely that our clustering manipulation induced interference between similar locations and that this caused the reduced precision that we observed in the clustered condition. Interestingly, these reductions in precision were perhaps only evident at longer retention intervals; the clustered and non-clustered conditions yielded numerically similar levels of precision after 0 and 3 hours (see Figure 2). Indeed, we predicted that there would be a more rapid loss of precision for clustered items relative to changes in accessibility (Hypothesis 5). Despite this, our *a priori* test for this interaction showed more evidence in favour of a null effect since the predicted difference was small and did not conform to our expectation of exponential changes across time (though our sensitivity threshold of $BF_{10} < 0.1$ was not reached). As such, the trade-off between accessibility and precision in this study does not appear to be modulated by retention interval.

As part of our planned exploratory analyses, we re-ran the main analyses but only included test trails where participants subjectively recognised the cue word as a previously studied item. The aim of this

was to produce measures of memory performance that reflect participants' ability to remember the word-location association when the word cue itself was subjectively recognised. After excluding trials that yielded no word recognition, we found that the effect of delay on accessibility scores was comparable to, if not larger than, the pre-registered effect. This suggests that the losses in memory accessibility were primarily driven by an inability for cue words to trigger associative retrieval rather than an inability to recognise the cue words themselves. Nonetheless, it is noteworthy that we are unable to determine whether word-location memories were not accessible due to a retrieval failure, or memory erasure (processes that Tulving termed failures of memory accessibility and availability, respectively[49–51]). A further possibility is that decreases in accessibility were driven by increases in misbinding, where the location of a different word is retrieved instead of the true word-location association[52].

The re-analyses involving only subjectively recognised cue words produced results that were largely the same as in the main analyses. However, it is noteworthy that this exploratory test yielded some evidence in favour of Hypothesis 5 (the interaction that was not originally supported). The reason for this divergent finding is not clear. Yet, there are two important differences between the pre-registered analyses and the exploratory analyses restricted by subjective memory judgments. First, the exploratory analysis only included data from a subset of test trials (i.e. trails where the cue word was subjectively recognised). It is possible that when a word was subjectively recognised but its location was not recalled, participants in later retention intervals relied on a form of spatial generalisation that yielded lower levels of precision. While possible, this account assumes that words which were not subjectively recognised, were also not subject to the same generalisation process, and this altered estimates of accessibility and precision in the main analysis. A second important difference relates to the number of participants who could be included in the exploratory analyses. Specifically, because restricting the number of test trials made participant exclusion more probable, the exploratory analyses involved disproportionally more participants with higher levels of retrieval confidence, particularly at longer retention intervals. Given this, the differing results may simply reflect a survivorship bias if subjective recognition confidence is correlated with memory precision.

Our planned exploratory analyses also examined time-dependent changes in the subjective memory judgments themselves. The number of responses indicating subjective word-location retrieval declined monotonically across retention intervals. However, there was one notable exception to this pattern; at 12 hours there were substantially fewer 'Word + location' and 'Word only' responses in comparison to the 6 hrs and 24 hrs intervals (see Figure 3). Importantly, estimates of accessibility and precision do not show this same non-monotonicity. The reason for this dissociation is not clear.

However, it is noteworthy that most participants in the 12 hrs condition (45/60 = 75%) ran their study phase in the morning and their test phase later that night (on average, night-time test sessions started at 22:47 local time). The remainder of these participants ran their study phase in the evening and their test phase early the next morning (average morning time session started at 07:31 local time). In contrast, participants in all other conditions tended to start both sessions during more regular working hours distributed through the day. As such, it possible that the reduction in subjective retrieval at 12 hours is attributable to psychological factors that fluctuate with the time-of-day, but importantly do not appear to affect participants' objective memory performance.

Our final set of exploratory analyses attempted to determine whether location responses became increasingly influenced by a spatial schema that represented approximate locations in the clustered condition. Kullback–Leibler divergence statistics indicated that the degree of pattern matching between participants' responses and the distribution of studied locations declined with time (see also[53]). Furthermore, this increasing divergence was strongly related to time-dependent changes in accessibility (I_p) rather than precision (I_k). This result would seem to be at odds with the theoretical position that generalised representations, perhaps supported by the neocortex, are more resilient to forgetting over time[54]. Our results suggest one of two possibilities. First, participants may not have been able to extract and/or use a generalised pattern when recalling the clustered associations. While evidence for time-dependent pattern extraction has been previously reported[38], it is possible that our stimulus set was not sufficiently structured to induce the use of a general pattern. Alternatively, it may have been possible that participants were indeed relying on a generalised pattern, yet the underlying representations supporting this were subject to the same time-dependent forgetting processes that affected non-clustered stimuli. This interpretation is supported by the finding that precision was overall lower in the clustered (relative to the non-clustered) condition, as it suggests that clustering leads to the development of a schema that confers less precise information. Importantly, retrieval-based generalisation mechanisms do predict that the loss of accessibility for specific items should be correlated with overall generalisation performance[55,56]. Thus, if participants are able to generalise in the clustered condition, our results are more in line with retrieval-based theories of generalisation. Further confirmatory research is required to investigate this possibility.

In sum, we have shown that forgetting distinct (non-overlapping) word-location associations predominantly involves losses in memory accessibility with negligible changes in memory precision. When memories do have similar features, and can potentially be represented by a general pattern, there is a strong performance trade-off resulting in increased accessibility but reduced precision across multiple retention intervals. However, this does not appear to confer a long-term increase in

the total amount of information that is maintained. Further, the trade-off does not appear to be significantly modulated by the retention interval between study and test. Our results are in line with theoretical models that predict generalisation performance is underpinned by retrieval-related accessibility for individual memory traces[55]. Additionally, our findings place constraints on computational models that make predictions about the nature of forgetting and generalisation, particularly in relation to the predicted robustness of generalised representations to forgetting[54,57].

Methods

Participants

Participants (native English-speaking, aged between 18 and 35 years) were recruited from Prolific (https://prolific.ac/). Prolific offers a web-based participant pool for behavioural scientists, manages participant payments, and ensures that individuals cannot participate in a given study more than once. All participants had either normal or corrected-to-normal vision (by self-report) and were compensated £7 for their time. The study was approved by a research ethics committee within the Department of Psychology at the University of York (ethical approval reference: 607).

Stimuli

A list of 200 common English nouns were used as stimuli (http://osf.io/8mzyc/). These belonged to one of 2 semantic categories: 100 manmade object nouns, and 100 natural object nouns. Words in each category were selected to be similar in length (mean difference: 0.020 characters; $d = 0.008$) and have a similar frequency in natural language (mean difference: 0.044; $d = 0.050$; as quantified by the Zipf scale in the Subtlex-UK database[40]). [Note: due to a minor coding error, the previous two effect sizes were mistakenly reported as being marginally larger in the original protocol registration, 0.011 and 0.063 respectively]. Additionally, we used a model of natural language word representations to ensure that the strength of semantic relationships between stimuli was similar in each category[41]. The word representations themselves were vectors in a 300-dimentional space and derived from a model that had been pre-trained on a set of web-based news articles containing approximately 100 billion words (see https://code.google.com/archive/p/word2vec). We took the Euclidian distance between vectors as a measure of semantic relatedness. This showed that there was only a trivial difference between the manmade and natural categories in terms of the mean semantic similarity between words ($d = 0.034$). [Note: due to a minor coding error this effect size was mistakenly reported as being 0.048 in the original protocol registration]. Nonetheless, a linear support vector machine was able to correctly classify 97% of the words as either manmade or natural using the vector representations alone. This suggests that the word categories were highly separable in semantic space. Finally, Kolmogorov–Smirnov tests showed that the distributions of word length, word frequency, and sematic relatedness did not substantially differ between the manmade and natural categories (each $D \leq 0.2$).

Procedure

Participants recruited from Prolific were directed to a secure website hosting the online experiment. An information sheet was shown detailing what the study involved including a description of the data that was collected and how it would be stored. At this time, participants were randomly allocated to one of 7 conditions; an immediate retrieval condition, which directly followed an initial study phase,

or a delayed retrieval condition (taking place either 3 hrs, 6 hrs, 12 hrs, 24 hrs, 48 hrs, or 96 hrs after the initial study phase). Before giving informed consent, participants were made aware of which condition they have been allocated to. They were told to revisit the experiment website within ±1 hour of their scheduled retrieval session to complete the task and obtain a full payment. A unique participant identifier was then provided by email which was used to start the retrieval session at the scheduled time. Participants were prevented from running any phase of the experiment on mobile devices such as handheld smartphones or tablets. Additionally, the task prevented participants from using devises with a screen resolution less than 600 x 600 pixels.

Study phase

During the study phase, a circular dial was visible in the centre of the screen. The task involved learning associations between different positions around this circle and specific words displayed on each trial (Figure 1A). All 200 word stimuli were presented at least once during the study phase. Words belonging to either the manmade or natural sematic categories were assigned to a 'clustered' condition. As such, they were associated with similar locations around the circle - randomly sampled from a von Mises distributions with a fixed width (k = 2.0), and a fixed mean (randomly chosen for each participant). All words belonging to the other semantic category were allocated to a 'non-clustered' condition. As such, they were associated with circular locations that had no consistent mean angle (von Mises concentration parameter, k < 0.05). The assignment of manmade/natural words to the clustered/non-clustered conditions was counterbalanced across participants.

Each study trial started with an indication of the circular position to be learned (location cue). A red cursor was drawn at a particular location along the circle's perimeter for 2 seconds (Figure 1A). Following this, the cursor was removed, and a study word was displayed onscreen for 4 seconds (word cue). Finally, with the word still visible, a red cursor was redrawn at a random location. Using the mouse/trackpad, participants were then required to verify that they had attended to the trial by repositioning the cursor at the cued location. This response window lasted 6 seconds for each trial and was followed by a 2-second inter-trail interval. If no response was made within the window, or if the response error is greater than 5°, the entire trial was repeated. Pilot data indicated that participants rarely repeated a given encoding trial more than 5 times. Nonetheless, to limit trial-to-trial variability in the encoding procedure, word cues that are repeated more than 5 times were excluded from the analyses. This study procedure is similar to that employed by previous investigations[17,18]. It is designed to ensure that participants attend to both the word and the location enabling an association to be learned between them.

Prior to starting the study phase, participants watched a short video demonstrating how the session was to progress, including instructions on how to make responses (video transcript available at http://osf.io/8mzyc/). These instructions emphasised that participants needed to remember the word-location associations as they were to be tested on them in the retrieval phase. As an aid to this, the video asked participants to imagine an object related to the cue word appearing just beside the cued location before responding to each study trial. Following the study phase, participants in the immediate retrieval condition completed the retrieval phase. Participants assigned to one of the delayed retrieval conditions were reminded of when they needed to revisit the experiment website.

Test phase

At test, participants were tasked with recalling each of the 200 word-location associations. As in the study phase, a circular dial was visible throughout. On each trial, a cue word was presented onscreen and, following a 1 second delay, a red cursor was drawn at a random location (Figure 1B). Participants then moved this cursor to the remembered location before making their response with a button press. Immediately after this, a prompt was shown asking participants to indicate whether they: (1) remembered both the word and its associated location ('Word + location'), (2) remembered the word but not its associated location ('Word only'), or (3) had forgotten encountering the word ('Neither'). Trails were separated by a 2 seconds inter-trail interval and a response window was imposed such that the next trail began automatically if both responses had not been made within 15 seconds (10 sec response window for the location judgement, 5 sec response window for the subjective memory judgement). We asked participants to be as accurate as possible, while ensuring that a response was made on every trial. They were also encouraged to make a best guess when entering location responses, even if they had no confidence in the accuracy of this response.

As in the study phase, all participants were shown a short video demonstrating how the retrieval session was to progress (video transcript available at http://osf.io/8mzyc/). After completing the retrieval phase, participants were then directed to a short questionnaire requesting a brief description of the strategy that they used when encoding and retrieving the word-locations associations. Participants were also be asked whether they had slept between the study and retrieval sessions and, if so, for how long. Following this, a debriefing sheet detailing the experimental hypotheses was provided. If participants in one of the delayed retrieval conditions attempted to start the test session more than one hour before their scheduled time slot, they were prevented from running the test and asked to return later. If participants missed their scheduled test session by over 1 hour, they were directed to a dedicated debriefing sheet informing them that they are unable to participate further. This further directed participants back to Prolific where they were reimbursed for the time spent

performing the study session (£3). Participants who returned to the experiment website after completing the test session were prevented from running the study and test phases a second time.

Recruitment protocol

An initial round of recruitment was run until we had collected 30 usable datasets per retention interval. At this point a statistical analysis of the data was performed and recruitment would have terminated if the Bayes factors relating to each of our *a priori* hypotheses were either greater than 10 (strong evidence in favour of an effect) or less than 0.1 (strong evidence in favour of no effect). If the Bayes factors did not show this level of sensitivity, data collection was to proceed in batches that added 10 usable datasets per retention interval. We planned to continue this until all Bayes factors had met the sensitivity threshold up to a maximum of 60 datasets per retention interval (420 complete datasets in total; maximum number dictated by resource constraints). Simulations based on our pilot data (see Supplementary Information) predicted that all Bayes factors were likely to reach the sensitivity threshold at a sample size of ~26 participants per retention interval.

Data analysis

Mixture model estimation

We simultaneously estimated retrieval probability (accessibility) and retrieval precision for individual participants using a probabilistic mixture model. First, we computed the replacement error of each response. This was given by the angular difference between a word's target location at study, and the retrieved location at test (see *Eq. S1*). For the mixture model, angular errors were assumed to be drawn from one of two distributions: (1) a circular uniform representing random guesses, and (2) a von Mises distribution representing the precision of memory retrieval. Each of these distributions has an associated prior probability; a statistic reflecting the overall proportion of responses belonging to that distribution. The prior for the von Mises distribution (donated p) encodes the rate of memory retrieval (i.e., retrieval probability; 'accessibility'). The von Mises distribution has two further parameters: a mean μ, and a dispersion statistic k (known as the 'concentration'). We fixed the value of μ to remain at zero, assuming that the average angular error of retrieved responses was always zero. The concentration parameter is analogous to the reciprocal of the variance; higher values of k indicate a narrower distribution. As such, k reflects the level of retrieval 'precision' and increases with better performance.

The parameters p (retrieval probability, 'accessibility') and k (memory precision) were estimated for clustered and non-clustered trials (separately) using an expectation-maximization (EM) algorithm (detailed in the Supplementary Methods, *Eq. S2–S6*; MATLAB functions available at http://osf.io/8mzyc/). This attempted to identify values of p and k that maximised the likelihood of

the observed data. The fit of the resulting mixture model was then compared to a reduced model that described all angular errors with a single uniform distribution (i.e., no mnemonic components). This comparison was made by calculating the difference in Bayesian information criterion statistics between models (ΔBIC, see *Eq. S7*). If the mixture model fitted the data substantially better than the reduced model (ΔBIC < -10), the parameters returned by the EM algorithm were accepted.

When the ΔBIC was greater than -10 (i.e., the mixture model provided a poor fit to the data relative to the reduced model) we used an alternative fitting procedure (see Supplementary Methods for details). The EM algorithm often fails to achieve a good fit when accessibility is low ($p \lesssim 0.2$; see Supplementary Methods). It was important to find a valid model fit to these datasets since merely excluding them would have resulted in a survivorship bias - overestimating a population's average performance because only the highest performing individuals are included. Here, the parameter p was systematically varied over a number of steps and k was estimated from the corresponding proportion of responses with the smallest angular error. This procedure can identify valid model fits as local minimum values of the likelihood function that are missed by the EM algorithm. If this produced a fit that was substantially better than the reduced model (as above; ΔBIC < -10), the parameters returned were accepted. However, if the alternative fitting procedure failed to return reliable estimates of both p and k for either the clustered or non-clustered condition, the participant's entire data set was excluded (exclusion criteria 6; see below).

Measures of memory-related information

While the model parameters p and k both reflect components of memory performance, these fundamentally different measures are not directly comparable. For instance, equivalent reductions in the values of p and k due to forgetting does not imply similar levels of forgetting in the form of accessibility and precision. We therefore use the differential entropy of angular errors to quantify the amount of mnemonic information that relates to each of these components. Entropy, denoted H, describes the uncertainty associated with observing a set of responses (i.e., angular errors) from a given distribution. If responses are highly uncertain (i.e., angular errors are widely dispersed around zero), entropy will be high. This indicates that the distribution generating responses (i.e., the word-location memories) conveys little positional information. The entropy of a von Mises distribution reflecting recollected responses is defined as follows:

$$H(k) = \int_{-\pi}^{\pi} f_{vm}(\theta|k) \cdot \log\left(\frac{1}{f_{vm}(\theta|k)}\right) d\theta \qquad \textit{Eq. 1}$$

Which simplifies to:

$$H(k) = \log\big(2\pi \cdot B_0(k)\big) - \frac{k \cdot B_1(k)}{B_0(k)} \qquad \textit{Eq. 2}$$

The term, $f_{vm}(\theta|k)$ denotes the probability density function for a von Mises distribution at angle θ, with a mean of 0 and concentration of k. The terms $B_0(k)$ and $B_1(k)$ refer to the modified Bessel function of the first kind with orders 0 and 1 (respectively), each evaluated at the point k. When k is zero, entropy is at a maxim ($H_{max} = \log(2\pi)$) and corresponds to that of the circular uniform distribution. This would imply that memory provides no positional information at all. In contrast, when k is large, (~17.5), entropy is near zero. This would suggest that responses are highly consistent with the learnt locations implying a large amount of mnemonic information. Given this, we subtract the entropy of recollected responses ($H(k)$) from the maximum possible entropy (H_{max}) to produce a measure of mnemonic information, denoted I_k:

$$I_k(k) = \log(2\pi) - H(k) \qquad \textit{Eq. 3}$$

This metric is 0 when precision is at a minimum and increases monotonically with more precise memories. However, increasing values of k to arbitrarily high levels results in only marginal increases in I_k. This reflects the fact that, beyond a certain point, increases in k produce only a small reduction in the angular span of the von Mises distribution.

Importantly, I_k is unweighted by the retrieval probability (p) and so does not consider the proportion of word-location pairs that are recalled. We therefore define a similar measure of information related to retrieval probability, I_p. As above, this is taken as the entropy (or uncertainty) associated with a given retrieval probability subtracted from H_{max}. The act of retrieving word-location associations rules-out random guesses (which are uniformly distributed). As such, the entropy associated with retrieval probability is taken as the uncertainty of random guessing (H_{max}) multiplied by the proportion of items that are not retrieved ($1 - p$). Subtracting this quantity from H_{max} yields a measure of mnemonic information (I_p) that is 0 when retrieval probability is minimal, and increases linearly to a value of $log(2\pi)$ when retrieval probability is at a maximum:

$$\begin{aligned} I_p(p) &= \log(2\pi) - (1-p) \cdot \log(2\pi) \\ &= p \cdot \log(2\pi) \end{aligned} \qquad \textit{Eq. 4}$$

As well as estimating the degree of mnemonic information associated with p and k separately, we also use a combined measure of mnemonic information to assess overall memory performance. This

measure, denoted I_t, reflects the total amount of information retained in memory given how many word-location pairs are retrieved, and the precision of the retrieved responses. It is computed by taking a sum of the entropies associated with memory recall and random guessing, weighted by retrieval probability, and subtracting the result from H_{max}:

$$I_t(p,k) = \log(2\pi) - (p \cdot H(k) + (1-p) \cdot \log(2\pi)) \qquad \textit{Eq. 5}$$

This measure also relates to I_p and I_k in the following way:

$$I_t = \frac{I_p \cdot I_k}{\log(2\pi)} \qquad \textit{Eq. 6}$$

Statistical modelling

Data from each participant were included in the analyses provided six criteria were met: (1) the participant successfully completed both study and test phases, (2) less than 20% of study trials were repeated more than 5 times (due to missed responses or poor replacement accuracy), (3) the number of retrieval trials that timed out did not exceed 30 within each condition, (4) the strategy description provided by participants at the end of testing does not suggest cheating or a lack of understanding regarding the task, (5) the dataset was uncorrupted and free of technical errors, and (6) a mixture model could be satisfactorily fit to the participants data as discussed in the methods and supplementary information. With regards to criterion 4, three independent raters (lab-members, including the 1st and 3rd authors), blind to the experimental conditions, reviewed the strategy descriptions and determined whether each participant had followed the task instructions appropriately. Individual participants were excluded if at least 2 of the 3 reviewers suspected cheating or a misunderstanding of the task.

Total information content of memory (I_t)

Hypotheses 1 and 2 concern the overall rate of forgetting (i.e., the loss mnemonic information measured by I_t), and whether clustering of locations for semantically related words improves overall memory performance (i.e., clustered vs non-clustered word-location pairs). To test these hypotheses, we specified a generalised-linear mixed-effects regression model (GLMM) to predict I_t within a 2x7 factorial structure (factor 1: clustering; factor 2: retention interval). Six binary coded predictors modelled the effect of each delayed retention interval (3 hrs, 6 hrs, 12 hrs, 24 hrs, 48 hrs, or 96 hrs) by contrasting them to the intercept term (representing immediate retrieval). Another binary predictor specified the effect of clustering by contrasting clustered vs non-clustered responses. Six further predictors coded the interaction between clustering and the delayed retention conditions.

In addition to the fixed effects predictors, a set of random effects parameters (2 per participant) were included to allow the intercept and clustering terms to freely vary across participants. All elements of the associated random effects covariance matrix were fully derived from the data. The model itself used a log-link function and was estimated via the maximum pseudolikelihood fitting method implemented in the MATLAB Statistics and Machine Learning Toolbox (MathWorks). Given that I_t is bounded by zero, the dispersion of responses was parametrised within the model using the gamma distribution. Pilot data (see Supplementary Information) revealed that this distribution provides a reasonable fit to the data and is better than all other commonly used distributions within the exponential family.

Table 2 lists each fixed-effects predictor and details the parameter contrast matrices that were used to test hypotheses 1 and 2. Hypothesis 1 examines whether there is a monotonic change in the total information metric across the 7 non-clustered retention intervals. To implement this, we ran a linear contrast that compared estimates of I_t across the intervals, weighted by the time difference between intervals. This required a contrast vector that, when multiplied with the delayed retention parameters (D1-D6), yields an effect size representing linear changes in these estimates over time (as in Table 2). Notably however, given that the GLMM uses a log link function, each parameter estimate reflects the log of I_t. This means that the linear contrast actually tests for exponential changes in I_t with respect to time. Exponential forgetting curves are known to provide a good fit to behaviour in both short-term and long-term memory experiments[42], as well as our pilot data (discussed in the supplementary information). Hypothesis 2 tests the main effect of clustering; i.e., whether there are overall differences in the total information metric between the clustered and non-clustered conditions. As such, this involved specifying a contrast vector that takes a weighted average across the 7 clustering predictors. Further details of how these contrast vectors were computed and applied to test our hypotheses are outlined in the Supplementary Methods.

Specific information content of memory (I_p and I_k)

Hypotheses 3, 4 and 5 concern differential rates of forgetting for clustered and non-clustered locations as measured by the two specific types of mnemonic information: I_p (accessibility) and I_k (precision). As above, we tested these hypotheses using a generalised-linear mixed-effects regression model (GLMM). The measures of mnemonic information, I_p and I_k, served as outcomes within this model, and the predictors constituted a 2x2x7 factorial structure (factor 1: memory type; factor 2: clustering; factor 3: retention interval).

As before, one binary predictor modelled the effect of clustering while a set of 6 dummy coded predictors specified the effect of each delayed retention interval. An additional binary predictor represented the difference between information types (I_p vs I_k). Finally, a set of 19 predictors modelled all interactions in the 3-factor structure. The model also included a set of random effects predictors (3 per participant) enabling the intercept, information type and clustering terms to freely vary across participants. All elements of the associated random effects covariance matrix were fully determined from the data. The model itself used a log link function, a gamma distribution to parameterise dispersion, and was estimated via the maximum pseudolikelihood fitting method.

Table 3 details how the fixed effect parameters of interest were contrasted in order to test hypotheses 3, 4, and 5. As in the previous GLMM, two of these involved testing for log-linear differences over time. Specifically, hypothesis 3 examined whether there was a two-way interaction between delay and information type (specifically in the non-clustered condition), while hypothesis 5 tested for a three-way interaction between delay, information type and clustering. As above, the contrast vectors for these hypotheses were designed to compare all parameter estimates of interest with each other, weighted by the time difference between retention intervals. Hypothesis 4 tested for an interaction between clustering and information type and therefore constituted a simple weighted average across the parameters coding for this effect. Further details of how these contrast vectors were computed and applied to test our hypotheses are outlined in the Supplementary Methods.

Bayesian inference

Each of our 5 *a priori* hypotheses were tested by computing Bayes factors in favour of a meaningful effect (denoted BF_{10}). Bayes factors greater than 10 indicate that, according to the data, there is at least 10 times more evidence in favour of the alternative hypothesis vs the null. Conversely, Bayes factors less than 0.1, indicate there is 10 times more evidence in favour of the null hypothesis over an alternative. When computing these statistics, we used a Cauchy distribution with a scale parameter of 0.555 to represent our prior uncertainty of standardised effect sizes (see *Eq. S10*). This scale factor is approximately the median effect size observed in our pilot study (see Supplementary Information). It was chosen such that the interval between the expected effect size and zero received a similar prior weight to the interval between the expected effect size and infinity. Full details of how these Bayes factors were computed are provided in the Bayesian inference section in the Supplementary Information. To complement each Bayes factor, standardised effect sizes are also reported. For completeness, we also report frequentist statistics, although these are not used to make inferences.

Exploratory analyses

As well as testing our pre-registered hypotheses, we also ran two additional planned exploratory analyses relating to the subjective memory judgements at the end of each retrieval trial. Given our lack of pilot data in relation to this aspect of the experiment, these are labelled as 'exploratory'. Continued data collection did not depend on the Bayes factors from these analyses as we had no *a priori* way of estimating how many participants would have been required to achieve sensitivity. Nonetheless, we report all *BFs*, standardised effect sizes, and frequentist inferential statistics related to these exploratory analyses.

First, we tested whether the subjective memory judgments provided at the end of each test trial suggested differential rates of forgetting for individual words (i.e., item memory) versus forgetting of work-location associations (i.e., associative memory). We specified a cumulative link mixed-effects regression model using the 'Ordinal' package in the *R* programming language. This accounted for relative changes in the proportion of test trials that were assigned either a 'Word + location', 'Word only', or 'Neither' response as a function of clustering and retention interval. The analysis therefore involved a 2x7 factorial structure with 3 responses categories. Random effects were modelled in the same way as in the total information GLMM discussed previously. The model used a logit link function and was estimated via the Gauss-Hermite fitting method. As this analysis involved subjective report data, it was not known *a priori* whether metacognitive response biases (e.g., a liberal tendency to respond 'Word only') would limit data quality and the conclusions that could have been drawn. Nonetheless, the model allowed us to assess whether changes in accessibility seen in the pre-registered analyses were primarily driven by forgetting of individual words (item memory) versus remembering the word but forgetting its associated location (associative memory).

Second, we assessed the relationship between word recognition as measured by subjective report, and mixture model estimates of accessibility and precision. Of specific interest was the extent to which losses in memory accessibility reflected either: (1) reduced accessibility for the cue word *per se*, or (2) failures to maintain the word-location association (in the presence of item memory for the word). To examine this, we re-ran the mixture models and GLMMs described above, but only included test trails where participants provided either a 'Word + location' or 'Word only' response. Excluding 'Neither' responses resulted in measures of memory accessibility (I_p) that reflect participants ability to remember the word-location association when the word cue itself was subjectively recognised. However, as this analysis was contingent on the proportion of words that receive either a 'Word + location' or 'Word only' response, it was again possible that metacognitive response biases would limit data quality. For instance, if 'Word + location' or 'Word only' responses were only made when

recognition strength was very high, only highly memorable trails would be included in the mixture model thereby potentially biasing estimates of accessibility and precision. Additionally, limiting the number of trails in the analysis is likely to have reduced the reliability of mixture model estimates in a way that does not uniformly affect each experimental condition.

Protocol registration

The Stage 1 protocol for this Registered Report was accepted in principle on 4th June 2019. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.c.4368464.v1.

Data availability

All anonymised behavioural data collected via the online task are freely available on the Open Science Framework (OSF) website (http://osf.io/8mzyc/).

Code availability

All HTML, PHP, and MATLAB scripts used to run the experimental task and analyse the data, are freely available on the OSF website (http://osf.io/8mzyc/).

References

1. Wagenaar, W. A. My memory: A study of autobiographical memory over six years. *Cogn. Psychol.* **18**, 225–252 (1986).

2. Ebbinghaus, H. Memory: A Contribution to Experimental Psychology. *Ann. Neurosci.* (1913) doi:10.5214/ans.0972.7531.200408.

3. McGeoch, J. A. Forgetting and the law of disuse. *Psychol. Rev.* (1932) doi:10.1037/h0069819.

4. Postman, L. Transfer, interference and forgetting. *Exp. Psychol.* 1019–1132 (1971).

5. Wixted, J. T. The Psychology and Neuroscience of Forgetting. *Annu. Rev. Psychol.* **55**, 235–269 (2004).

6. Sadeh, T., Ozubko, J. D., Winocur, G. & Moscovitch, M. How we forget may depend on how we remember. *Trends Cogn. Sci.* **18**, 26–36 (2014).

7. Richards, B. A. & Frankland, P. W. The Persistence and Transience of Memory. *Neuron* **94**, 1071–1084 (2017).

8. Hardt, O., Nader, K. & Nadel, L. Decay happens: The role of active forgetting in memory. *Trends Cogn. Sci.* **17**, 111–120 (2013).

9. Reyna, V. F. & Brainerd, C. J. Fuzzy-Trace Theory: Interim Theory Synthesis. *Learn. Individ. Differ.* **7**, 1–75 (1995).

10. Nadel, L. & Moscovitch, M. Memory consolidation , retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).

11. Winocur, G. & Moscovitch, M. Memory transformation and systems consolidation. *J. Int. Neuropsychol. Soc.* **17**, 766–780 (2011).

12. Sekeres, M. J., Winocur, G. & Moscovitch, M. The hippocampus and related neocortical structures in memory transformation. *Neurosci. Lett.* **680**, 39–53 (2018).

13. Murphy, G. L. & Shapiro, A. M. Forgetting of verbatim information in discourse. **22**, 84–94 (1994).

14. Kintsch, W., Welsch, D., Schmalhofer, F. & Zimny, S. Sentence Memory: A Theoretical. **159**, 133–159 (1990).

15. Sekeres, M. J. *et al.* Recovering and preventing loss of detailed memory: Differential rates of forgetting for detail types in episodic memory. *Learn. Mem.* **23**, 72–82 (2016).

16. Furman, O., Hasson, U., Davachi, L., Dorfman, N. & Dudai, Y. They saw a movie: Long-term memory for an extended audiovisual narrative. *Learn. Mem.* **14**, 457–467 (2007).

17. Harlow, I. M. & Donaldson, D. I. Source accuracy data reveal the thresholded nature of human episodic memory. *Psychon. Bull. Rev.* **20**, 318–325 (2013).

18. Harlow, I. M. & Yonelinas, A. P. Distinguishing between the success and precision of recollection. *Memory* (2016) doi:10.1080/09658211.2014.988162.

19. Richter, F. R., Cooper, R. A., Bays, P. M. & Simons, J. S. Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *Elife* **5**, 1–18 (2016).

20. Nilakantan, A. S., Bridge, D. J., VanHaerents, S. & Voss, J. L. Distinguishing the precision of spatial recollection from its success: Evidence from healthy aging and unilateral mesial temporal lobe resection. *Neuropsychologia* **119**, 101–106 (2018).

21. Nilakantan, A. S., Bridge, D. J., Gagnon, E. P., VanHaerents, S. A. & Voss, J. L. Stimulation of the Posterior Cortical-Hippocampal Network Enhances Precision of Memory Recollection. *Curr. Biol.* **27**, 465–470 (2017).

22. Schurgin, M. W., Wixted, J. T. & Brady, T. F. Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength. (2018).

23. Sun, S. Z. *et al.* Erasing and blurring memories: The differential impact of interference on separate aspects of forgetting. *J. Exp. Psychol. Gen.* **146**, 1606–1630 (2017).

24. Luck, S., Vogel, J. & Edward, K. The Capacity of Visual Working Memory for Features and Conjuctions. *Nature* **390**, 279–281 (1997).

25. Bays, P. M., Catalao, R. F. G. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* **9**, 7–7 (2009).

26. Murray, J. G., Howie, C. A. & Donaldson, D. I. The neural mechanism underlying recollection is sensitive to the quality of episodic memory: Event related potentials reveal a some-or-none threshold. *Neuroimage* **120**, 298–308 (2015).

27. Cai, D., Kleeman, R. & Majda, A. A Mathematical Framework for Quantifying Predictability Through Relative Entropy. *Methods Appl. Anal.* **9**, 425–444 (2002).

28. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

29. Verdugo Lazo, A. C. G. & Rathie, P. N. On the Entropy of Continuous Probability Distributions. *IEEE Trans. Inf. Theory* **24**, 120–122 (1978).

30. Bartlett, F. F. C. Remembering: An experimental and social study. *Cambridge: Cambridge University* (1932).

31. Ghosh, V. E. & Gilboa, A. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* **53**, 104–114 (2014).

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

33. Mcclelland, J. L., Mcnaughton, B. L. & Reilly, R. C. O. McClelland PsycRev 1995 HPC-Neocortex.pdf. **102**, 419–457 (1995).

34. Kan, I. P., Alexander, M. P. & Verfaellie, M. learning in amnesia. **21**, 938–944 (2009).

35. Arpit, D. *et al.* A Closer Look at Memorization in Deep Networks. (2017).

36. Richter, F. R., Bays, P. M., Jeyarathnarajah, P. & Simons, J. S. Flexible updating of dynamic

knowledge structures. *Sci. Rep.* **9**, 2272 (2019).

37. Brady, T. F., Schacter, D. L. & Alvarez, G. A. The Adaptive Nature of False Memories is Revealed by Gist-based Distortion of True Memories. *PsyArXiv* (2018) doi:10.1167/15.12.948.The.

38. Richards, B. A. *et al.* Patterns across multiple memories are identified over time. *Nat. Neurosci.* **17**, 981–986 (2014).

39. Mack, M. L., Preston, A. R. & Love, B. C. Decoding the brain's algorithm for categorization from its neural implementation. *Curr. Biol.* **23**, 2023–2027 (2013).

40. van Heuven, W. J. B., Mandera, P., Keuleers, E. & Brysbaert, M. SUBTLEX-UK: A new and improved word frequency database for British English. *Q. J. Exp. Psychol.* **67**, 1176–1190 (2014).

41. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. 1–12 (2013) doi:10.1162/153244303322533223.

42. Rubin, D. C. & Wenzel, A. E. One hundred years of forgetting: A quantitative description of retention. *Psychol. Rev.* **103**, 734–760 (1996).

43. Diamond, N., Armson, M. & Levine Brian. The truth is out there: Accuracy and detail in recall of verifiable real-world events. *PsyArXiv* (2019).

44. Joensen, B. H. *et al.* Journal of Experimental Psychology : General Episodic Events United We Fall : All-or-None Forgetting of Complex Episodic Events. (2019).

45. Davis, R. L. & Zhong, Y. The Biology of Forgetting—A Perspective. *Neuron* **95**, 490–503 (2017).

46. Frankland, P. W., Köhler, S. & Josselyn, S. A. Hippocampal neurogenesis and forgetting. *Trends Neurosci.* **36**, 497–503 (2013).

47. Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A. & Tonegawa, S. Engram cells retain memory under retrograde amnesia. *Science (80-.).* **348**, 1007–1013 (2015).

48. Roy, D. S., Muralidhar, S., Smith, L. M. & Tonegawa, S. Silent memory engrams as the basis for retrograde amnesia. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9972–E9979 (2017).

49. Tulving, E. Ecphoric processes in episodic memory. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **302**, 361–371 (1983).

50. Tulving, E. *Elements of episodic memory*. (1983).

51. Frankland, P. W., Josselyn, S. A. & Köhler, S. The neurobiological foundation of memory retrieval. *Nature Neuroscience* vol. 22 1576–1585 (2019).

52. Pertzov, Y. *et al.* Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain* **136**, 2474–2485 (2013).

53. Tompary, A., Zhou, W. & Davachi, L. Schematic memories develop quickly, but are not expressed unless necessary. *PsyArXiv* (2020).

54. Kumaran, D., Hassabis, D. & McClelland, J. L. What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).

55. Kumaran, D. & McClelland, J. L. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychol. Rev.* **119**, 573–616 (2012).

56. Kumaran, D. What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Front. Hum. Neurosci.* **6**, 1–11 (2012).

57. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, (2017).

Acknowledgments

We thank Sara Mod, Lamia Begic, Tabitha Houldridge, and Thomas Maltby for help with collecting the lab-based pilot data. AJH is funded by the Wellcome Trust (204277/Z/16/Z) and ESRC (ES/R007454/1). BR is funded by a Learning in Machines and Brains Fellowship from the Canadian Institute for Advanced Research and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-04947). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

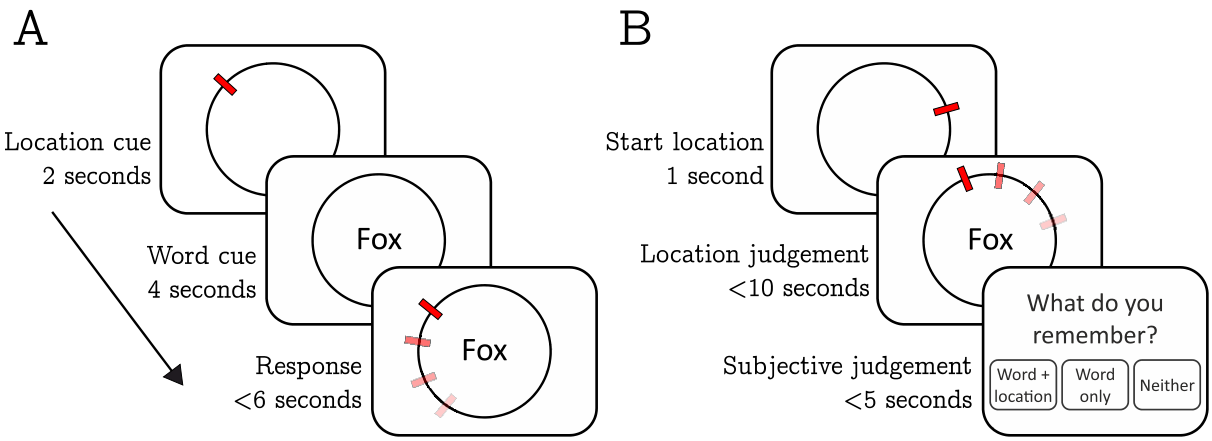
SCB, BAR, and AJH contributed to research design and provided input on the manuscript. SCB and AJH wrote the manuscript and developed the analysis pipeline. SCB coded the experimental tasks, derived the experimental metrics, and implemented the statistical analyses.

Competing interests

The authors declare no competing interests.

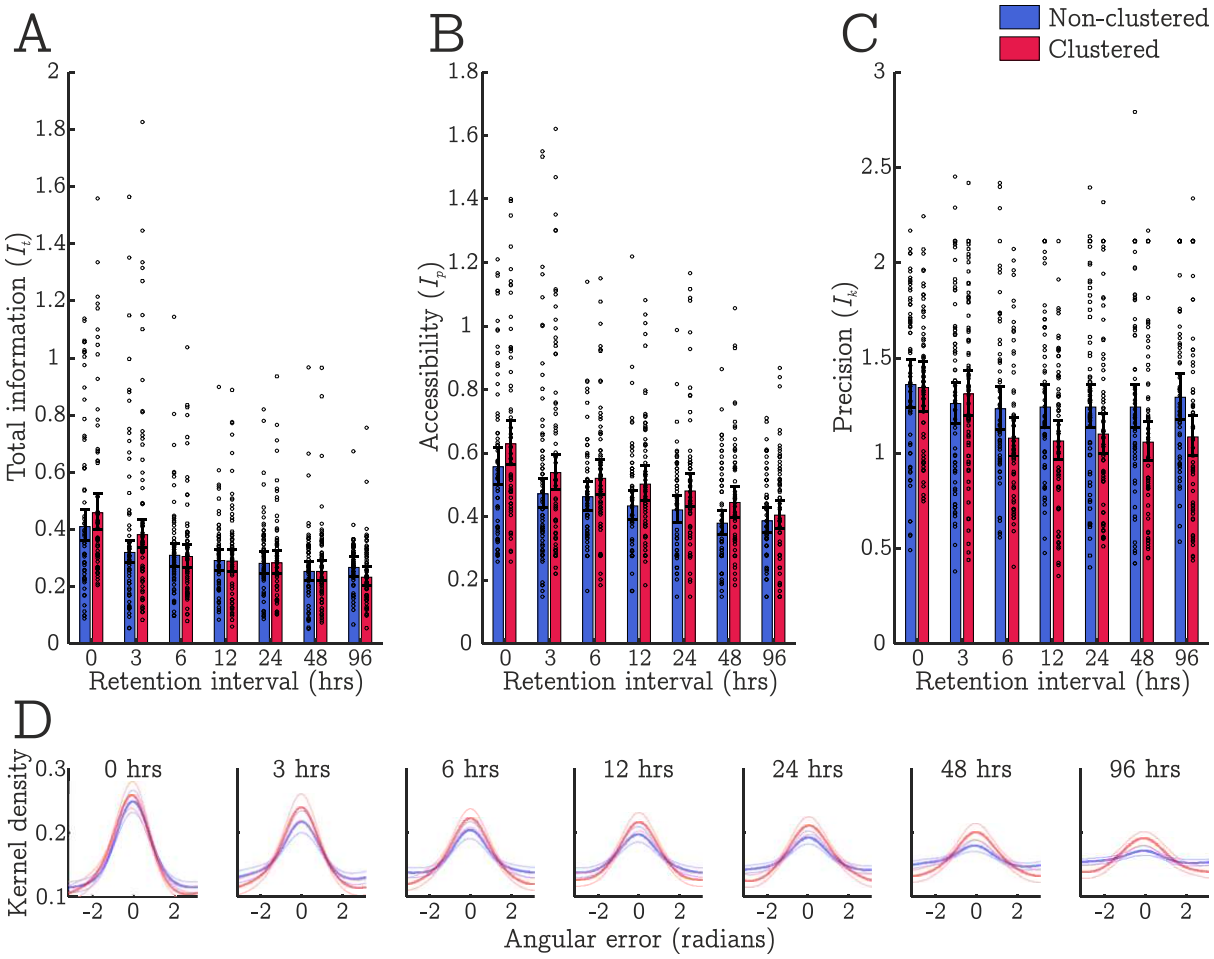
Figures

Figure 1



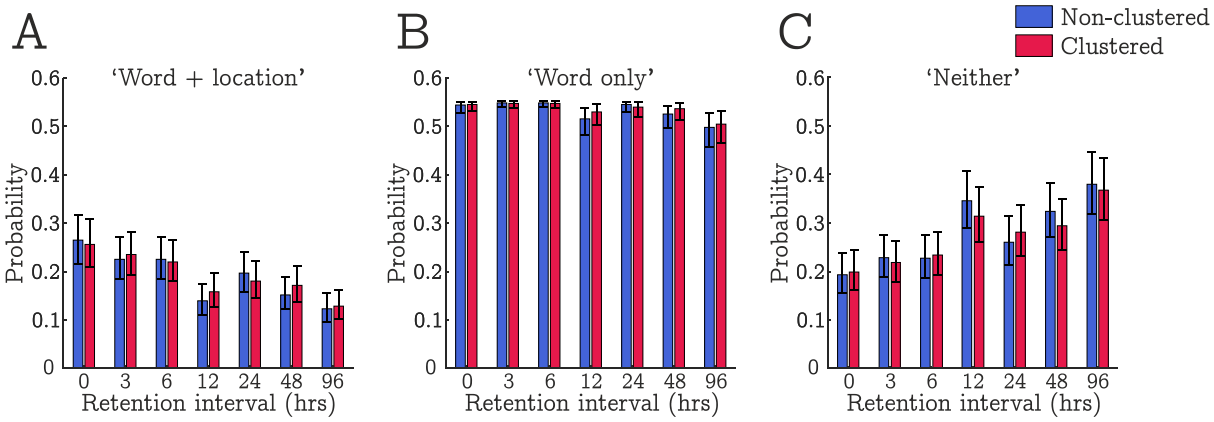
Schematic of the experimental procedure. *A:* Structure of a Study trial. A location cursor was presented for 2 secs, followed by the word cue. The cursor then reappeared in a randomly chosen location and the participant was required to move it back to the recently shown location (to within 5°). *B*: Structure of a Test trial. A location cursor was presented in a random location for 1 sec, followed by a word previously shown at study. The participant was required to move the cursor to the remembered location associated with that word (location judgement; 10 sec response window). Following this, participants were asked to indicate whether they remembered both the word and its associated location, the word alone, or neither of the two (subjective judgement; 5 sec response window).

Figure 2



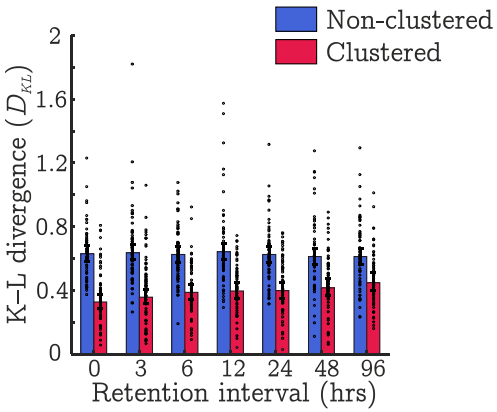
Memory performance by condition. *Top row:* Means (and 95% confidence intervals) for each measure of mnemonic information plotted by retention interval and clustering condition; (*A*)**.** Total information content, I_t. (*B*)**.** Accessibility information content, I_p. (*C*)**.** Precision information content, I_k. Individual datapoints represent participant scores. *Bottom row, D:* Kernel density estimates plotting the average distributions of angular errors in each condition with 95% confidence bounds. Each panel showed the results from a different retention interval with 0 hrs on the extreme left and 96 hrs on the extreme right. Blue curves denote estimates from non-clustered trials while red curves denote estimates from clustered trials.

Figure 3



Model-derived probability estimates for each type of the subjective memory judgment. From left to right, each panel plots the proportion of 'Word + location' *(A)*, 'Word only' *(B)*, and 'Neither' responses *(C)*. Note that these estimates were returned by a cumulative link mixed-effects regression model. This was estimated from a large list of categorical responses coding the subjective judgment that was made on each test trial. As the outcome was an ordinal variable, individual datapoints are not plotted.

Figure 4



Kullback–Leibler divergence (D_{KL}) by condition. Mean D_{KL} estimates as a function of retention interval and clustering condition. Error bars represent 95% confidence intervals and individual datapoints depict participant scores. Note that lower D_{KL} scores indicate a closer correspondence between the absolute position of location responses and the experimentally imposed spatial pattern in the clustered condition.

Tables

Table 1

Effect sizes, Bayes factors, and frequentist statistics for each preregistered hypothesis. 95% confidence intervals are indicated in square brackets.

	Cohen's D	*BF_{10}*	*T statistic*	*95% CI*	*P value*
Hypothesis 1 Change in total information across delay in the non-clustered condition.	0.314	117	3.787 *(d.f. = 848)*	[0.127, 0.402]	< .001
Hypothesis 2 Difference in total information between clustered and non-clustered condition.	0.021	0.054	0.743 *(d.f. = 848)*	[-0.030, 0.066]	.458
Hypothesis 3 The effect of delay differs between accessibility and precision in the non-clustered condition.	0.275	35.1	3.449 *(d.f. = 1696)*	[0.113, 0.410]	< .001
Hypothesis 4 Clustering differentially effects accessibility vs precision.	0.141	5.98×10^6	6.179 *(d.f. = 1696)*	[0.144, 0.278]	< .001
Hypothesis 5 Clustering changes the difference between accessibility and precision as a function of delay.	0.074	0.188	0.665 *(d.f. = 1696)*	[-0.132, 0.267]	.506

Table 2

Parameter contrast matrices for hypotheses 1 and 2 tested by the GLMM of total information (I_t).

	Regressor name												
	Clustered (C)	**Delay 1 (D1)**	**Delay 2 (D2)**	**Delay 3 (D3)**	**Delay 4 (D4)**	**Delay 5 (D5)**	**Delay 6 (D6)**	**C * D1**	**C * D2**	**C * D3**	**C * D4**	**C * D5**	**C * D6**
Hypothesis 1 Change in total information across delay in the non-clustered condition.	**0**	**.299**	**.261**	**.187**	**.037**	**-.261**	**-.859**	**0**	**0**	**0**	**0**	**0**	**0**
Hypothesis 2 Difference in total information between clustered and non-clustered condition.	**.944**	**0**	**0**	**0**	**0**	**0**	**0**	**.135**	**.135**	**.135**	**.135**	**.135**	**.135**

Table 3

Parameter contrast matrices for hypotheses 3-5 for the GLMM for specific information content (I_p, I_k). Note, not all model parameters are listed; the model additionally includes parameters accounting for the non-interacting effects of information-type, clustering and delay. T = information type [I_p vs I_k]; C = clustering; D = delay condition.

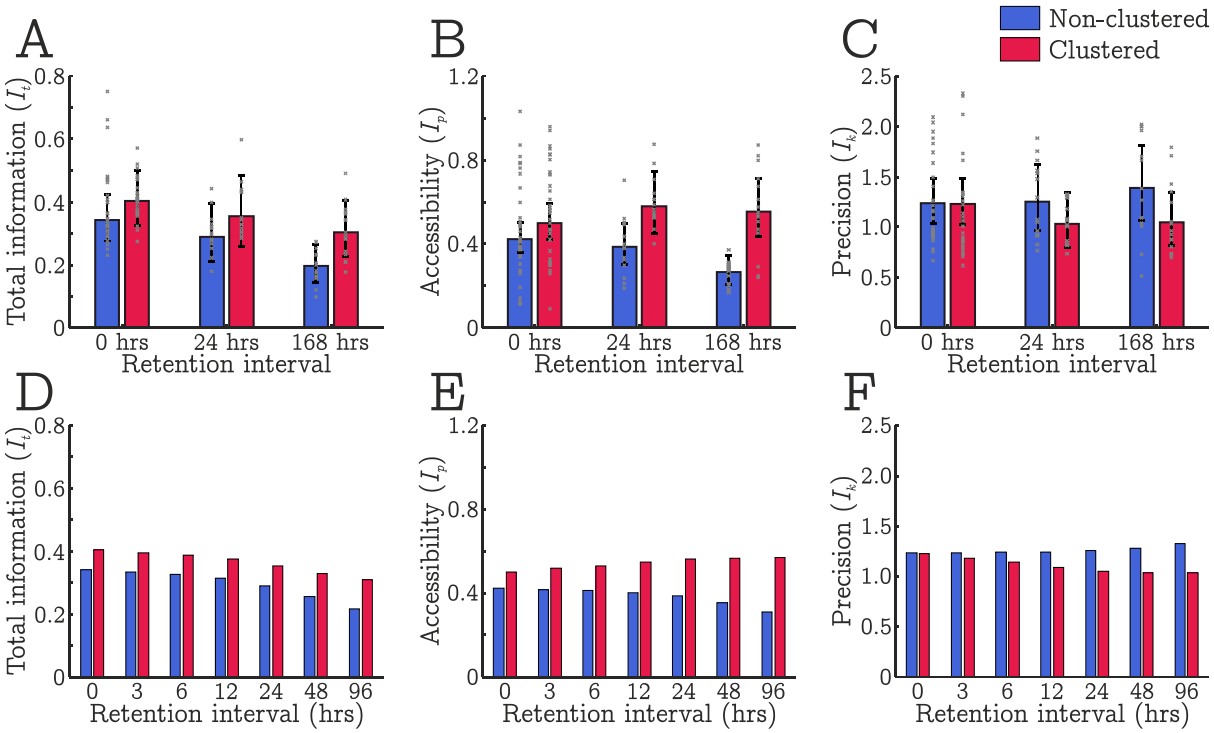
	Regressor name												
	Info-type (T) * C	**T * D1**	**T * D2**	**T * D3**	**T * D4**	**T * D5**	**T * D6**	**C * T * D1**	**C * T * D2**	**C * T * D3**	**C * T * D4**	**C * T * D5**	**C * T * D6**
Hypothesis 3 The effect of delay will differ between accessibility and precision in the non-clustered condition.	**0**	**.299**	**.261**	**.187**	**.037**	**-.261**	**-.859**	**0**	**0**	**0**	**0**	**0**	**0**
Hypothesis 4 Clustering differentially effects accessibility vs precision.	**.944**	**0**	**0**	**0**	**0**	**0**	**0**	**.135**	**.135**	**.135**	**.135**	**.135**	**.135**
Hypothesis 5 Clustering changes the difference between accessibility and precision as a function of delay.	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**.299**	**.261**	**.187**	**.037**	**-.261**	**-.859**

Supplementary Information

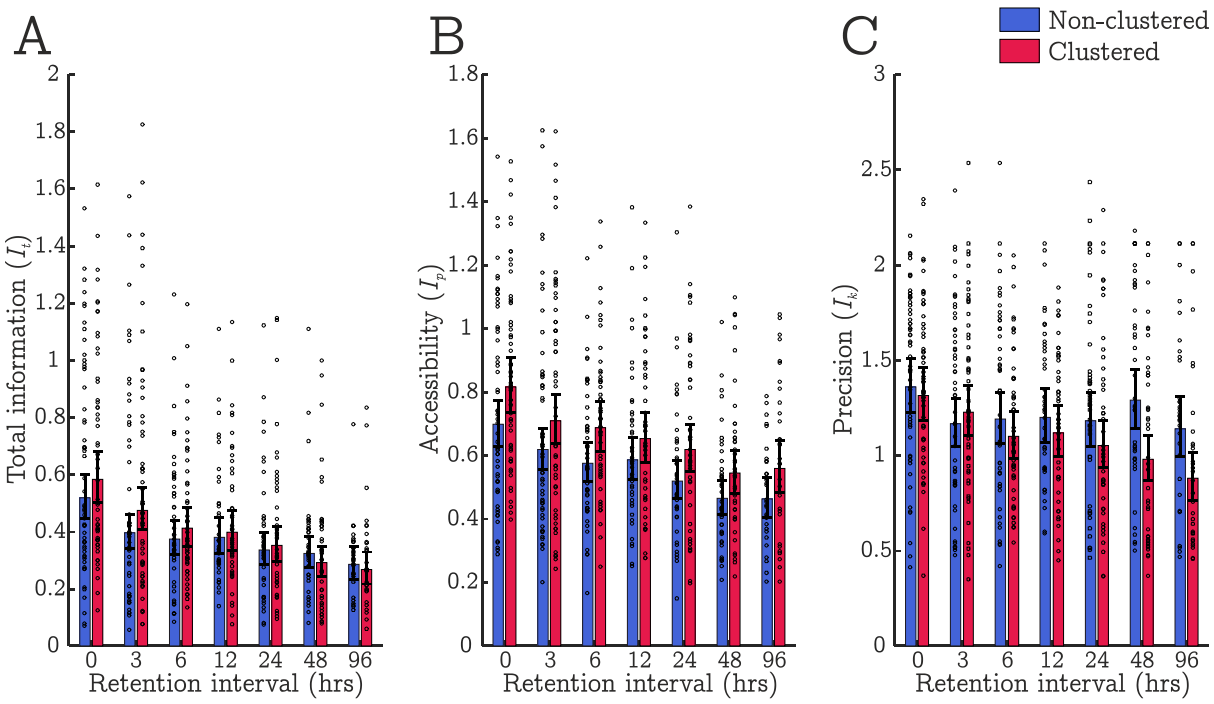
"Dissociating memory accessibility and precision in forgetting"

Sam C Berens, Blake A Richards, & Aidan J Horner

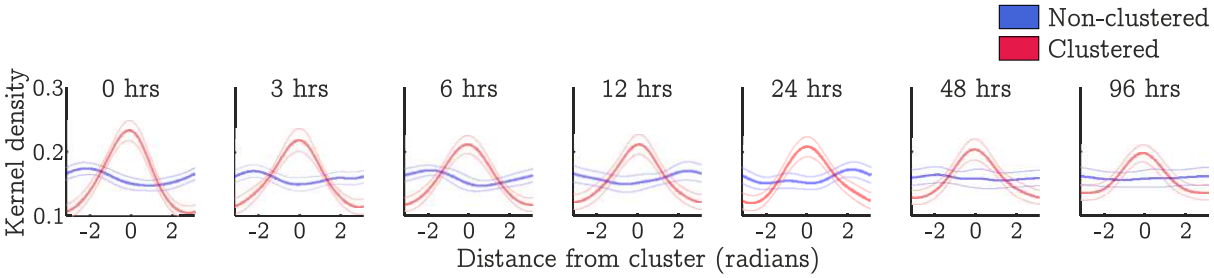
Supplementary Figures



Supplementary Figure 1. Results of pilot study and simulated data. ***Top row:*** Means (and 95% confidence intervals) in the pilot data for each measure of mnemonic information plotted by retention interval and clustering condition; ***A.*** Total information content, I_t. ***B.*** Accessibility information content, I_p. ***C.*** Precision information content, I_k. Individual datapoints represent participant scores after controlling for random intercepts (n=73 in both the clustered and non-clustered conditions within each panel). ***Bottom row:*** Mean estimates for the clustered and non-clustered conditions at each of the 7 retention internals in the main experiment; ***D.*** Total information content (I_t) ***E.*** Accessibility (I_p) and ***F.*** Precision (I_k). Estimates are based on fitting the pilot data to the exponential model of forgetting in *Eq S12*.



Supplementary Figure 2. Memory performance after excluding trials with no word recognition. Means (and 95% confidence intervals) for each measure of mnemonic information plotted by retention interval and clustering condition; ***A.*** Total information content, I_t. ***B.*** Accessibility information content, I_p. ***C.*** Precision information content, I_k. Individual datapoints represent participant scores.



Supplementary Figure 3. Spatial patterns of participant responses. Kernel density estimates plotting the average distributions of replaced locations in each condition with 95% confidence bounds. Importantly, these estimates reflect the absolute position of responses relative to centre of the experimentally imposed cluster. Each panel showed the results from a different retention interval with 0 hrs on the extreme left and 96 hrs on the extreme right. Blue curves denote estimates from non-clustered trials while red curves denote estimates from clustered trials.

Supplementary Tables

Supplementary Table 1. Statistical analysis of pilot data. Standardised effect sizes and Bayes factors for hypotheses 1-5. As effect sizes were uncertain *a priori*, the Bayes factors were calculated using a Cauchy scale factor of $\sqrt{0.5}$.

	Cohen's D	***BF_{10}***
Hypothesis 1	0.486	6.573
Hypothesis 2	0.545	7.238×10^{2}
Hypothesis 3	0.523	23.40
Hypothesis 4	0.596	5.116×10^{5}
Hypothesis 5	0.645	10.93

Supplementary Table 2. Means (and standard deviations) of each outcome measure in the final sample, broken down by condition.

	Retention interval	p	k	I_p	I_k	I_t
Non- clustered	**0 hrs**	.328 *(.150)*	10.733 *(7.645)*	0.603 *(0.275)*	1.413 *(0.441)*	0.490 *(0.314)*
	3 hrs	.285 *(.160)*	10.609 *(10.833)*	0.524 *(0.295)*	1.328 *(0.500)*	0.390 *(0.298)*
	6 hrs	.262 *(.092)*	9.396 *(10.953)*	0.481 *(0.169)*	1.281 *(0.442)*	0.336 *(0.175)*
	12 hrs	.249 *(.100)*	8.772 *(7.871)*	0.458 *(0.184)*	1.288 *(0.419)*	0.311 *(0.141)*
	24 hrs	.240 *(.086)*	9.953 *(9.947)*	0.441 *(0.157)*	1.307 *(0.484)*	0.306 *(0.146)*
	48 hrs	.218 *(.087)*	11.550 *(16.249)*	0.401 *(0.160)*	1.320 *(0.533)*	0.279 *(0.148)*
	96 hrs	.220 *(.075)*	9.753 *(8.816)*	0.404 *(0.137)*	1.341 *(0.423)*	0.278 *(0.088)*
Clustered	**0 hrs**	.370 *(.163)*	9.637 *(7.197)*	0.680 *(0.300)*	1.387 *(0.370)*	0.528 *(0.329)*
	3 hrs	.331 *(.185)*	10.547 *(9.503)*	0.608 *(0.341)*	1.373 *(0.456)*	0.467 *(0.359)*
	6 hrs	.299 *(.114)*	6.436 *(5.776)*	0.549 *(0.209)*	1.127 *(0.396)*	0.341 *(0.197)*
	12 hrs	.289 *(.113)*	6.261 *(5.036)*	0.532 *(0.207)*	1.122 *(0.416)*	0.327 *(0.184)*
	24 hrs	.281 *(.126)*	8.139 *(9.092)*	0.517 *(0.232)*	1.171 *(0.486)*	0.315 *(0.171)*
	48 hrs	.256 *(.101)*	7.315 *(7.938)*	0.470 *(0.185)*	1.127 *(0.473)*	0.283 *(0.164)*
	96 hrs	.237 *(.095)*	8.777 *(10.170)*	0.436 *(0.175)*	1.174 *(0.526)*	0.253 *(0.116)*

Supplementary Table 3. Results of planned exploratory analysis. Effect sizes, Bayes factors, and frequentist statistics for each hypothesis after excluding test trials that received a 'Neither' response. 95% confidence intervals are indicated in square brackets.

	Cohen's D	*BF_{10}*	*T statistic*	*95% CI*	*P value*
Hypothesis 1	0.458	280	3.985 *(d.f. = 624)*	[0.201, 0.590]	< .001
Hypothesis 2	0.055	0.157	1.467 *(d.f. = 624)*	[-0.014, 0.100]	.143
Hypothesis 3	0.322	7.699	2.886 *(d.f. = 1248)*	[0.087, 0.458]	.004
Hypothesis 4	0.221	2.44×10^{10}	7.377 *(d.f. = 1248)*	[0.186, 0.321]	< .001
Hypothesis 5	0.406	4.472	2.613 *(d.f. = 1248)*	[0.074, 0.518]	.009

Supplementary Methods

Mixture model estimation

We first computed the angular error of each response in radians (denoted x_i). This is taken as the angular difference between the target location seen at study (θ), and the retrieved location entered at test ($\hat{\theta}$).

$$x_i = (\hat{\theta_i} - \theta_i) \bmod 2\pi \qquad \textit{Eq. S1}$$

Given these errors, estimation via the EM algorithm started by first assigning arbitrary random values to the parameters being estimated. The algorithm then progressed in two steps (an E-step and an M-step) that were repeated in sequence across multiple iterations. During the E-step, we computed a set of weightings (w_i) representing the probability that individual responses were based on memory retrieval (von Mises distributed errors). These weightings were dependent on the angular error x_i as well as the two model parameters p and k.

$$w_i(x_i \mid p, k) = \frac{p \cdot f_{vm}(x_i|k)}{p \cdot f_{vm}(x_i|k) + (1-p) \cdot (2\pi)^{-1}} \qquad \textit{Eq. S2}$$

The quantity $f_{vm}(x_i|k)$ denotes the probability density function for a von Mises distribution at angle x_i with a mean of 0 and concentration of k, see[1]). Note that term $(2\pi)^{-1}$ reflects the probability density function of the circular uniform distribution for any value of x_i. Given the weighing w_i for each response, we computed new values for each model parameter (the M-step). The parameter p was computed as follows:

$$p = \sum_{i=1}^{n} \frac{w_i}{n} \qquad \textit{Eq. S3}$$

To re-estimate the parameter k, we first computed the population resultant vector (r), the average of all response errors weighted by the probability that they belong to the von Mises distribution (w_i).

$$r = \text{real}\left(\frac{\sum_{i=1}^{n}(w_i \cdot \exp(j \cdot x_i))}{\sum_{i=1}^{n} w_i}\right) \qquad \textit{Eq. S4}$$

Where j denotes the imaginary unit. The statistic r was then converted into the concentration parameter k, using an approximation provided by Fisher[1].

$$k = \begin{cases} 2r + r^3 + \frac{5r^5}{6}, & r < 0.53 \\ -0.4 + 1.39r + \frac{0.43}{1-r}, & 0.53 \leq r < 0.85 \\ \frac{1}{3r - 4r^2 + r^3}, & r \leq 0.85 \end{cases} \qquad \textit{Eq. S5}$$

This approximation of k is known to be heavily biased when it is based on fewer than 15 data points (i.e., when p is low[2]). As such, in a final step, we applied the following correction to estimates of k as suggested by Best and Fisher[2]:

$$k^* = \begin{cases} k, & n \cdot p > 15 \\ \begin{cases} \frac{(n \cdot p - 1)^3 \cdot k}{n \cdot p(n^2 \cdot p^2 + 1)}, & k \geq 2 \\ \max(k - \frac{2}{n \cdot p \cdot k}, 0), & k < 2 \end{cases}, & n \cdot p \leq 15 \end{cases} \qquad \textit{Eq. S6}$$

Where n is the number of word-location trails (in this case 100), and $k^*$is the adjusted estimate of k.

These estimation steps repeated until the negative log-likelihood (NLL) of the model (i.e., the goodness-of-fit), converged to a stable value. The EM algorithm is sensitive to the starting values assigned to each parameter and can converge at local minimum values of the NLL function. As such, each estimation was run with 17 unique starting points using 17 linearly spaced values of p and a starting value of k = 2 each time. These starting points were found to yield the most accurate results when analysing pilot data. The iteration with the lowest NLL was then selected as the final model.

Assessing model fit

In cases where a participant's retrieval probability was low ($p \lesssim 0.2$), the EM algorithm may have failed to converge or may have incorrectly fit a wide von Mises distribution indistinguishable from a uniform ($k \approx 0.1$). This latter case results in inflated estimates of retrieval probability since the similarly shaped uniform and von Mises distributions will provide equal weightings to all data points (i.e., $w \approx 0.5$). This pathological case can be identified by comparing complexity-adjusted measures of goodness-of-fit between the final mixture model and a reduced model that describes all data points with a single uniform distribution. Here, we used the difference in the Bayesian information criterion (denoted ΔBIC) to make this comparison[3]. Given that the mixture model has 2 free parameters, p and k, and the reduced model has no free parameters, the ΔBIC was computed as follows:

$$\Delta BIC = 2 \cdot \left(\log(n) - \log\left(\hat{L}_m\right) + \log(L_u)\right) \qquad \textit{Eq. S7}$$

The term, $\log(\hat{L}_m)$ denotes the log-likelihood of the mixture model, and $\log(L_u)$ denotes the log-likelihood of the reduced model, in this case, a constant value of $-n \cdot \log(2\pi)$. As such, lower (more negative) values of ΔBIC indicate that the mixture model provides a better fit to the data than the reduced model after accounting for the additional complexity. We took ΔBIC values of -10 or below to indicate that the model had converged properly, and the parameters were reliable. This threshold is often used to represent strong evidence for the more complex model[4] and we found it to reliably distinguish pathological and valid solutions in our pilot data.

Alternative fitting procedure

In cases were the EM algorithm returned a ΔBIC greater than the -10 threshold, or failed to converge altogether, we attempted to identify a valid fit via an alternative search procedure. At first, this involved explicitly varying the retrieval probability (p) over a number of steps (from p = 0.02 to 0.3; 2-30 words) before estimating k and the NLL (as above) from the $p \cdot n$ most accurate responses (a so-called 'hard-clustering' approach). This often identifies local minimum values of the NLL function that are missed by the EM algorithm. We accepted mixture model estimates identified in this way as long as the corresponding ΔBIC statistic was below our -10 threshold. Importantly however, this procedure often returns estimates of k that are not reliable when based on fewer than 8 responses, even after applying the correction expressed in *Eq. S6* (singularities can result, causing k to become arbitrary large). We therefore excluded data from participants when this was the case. If no mixture model could be fit to a participant's data such that the ΔBIC statistic was less than -10, the participant was excluded from further statistical analyses.

Linear contrasts

Hypotheses 1, 3, and 5, involved testing for differences or interactions across the 7 retention intervals. As stated in the main text, this entails contrasts that are sensitive to linear changes in the GLMM parameter estimates over time. To implement this, we specified a 1-by-6 contrast vector, $H = [h_1, h_2, h_3, h_4, h_5, h_6]$, that evaluated differences between pairs of parameters, and weighted these differences by the time between retention intervals. Each element of H was given by the following expression:

$$h_i = \sum_{a=1}^{6} \left(T_a - \frac{7}{6} \cdot T_i \right) \qquad \textit{Eq. S8}$$

Where, T is a 6D vector encoding the retention time (in hours) of each delayed interval: $T = [3, 6, 12, 24, 48, 96]$. The scaling factor of $7/6$ ensured that each delayed retention interval (i) was compared to the immediate retrieval condition (represented by the intercept term) as well as every

other delayed condition. The resulting vector was then scaled to have a unit length by dividing each element by the overall magnitude. This produced a set of contrast weights that linearly decreased as a function of retention time. Consequently, performing a matrix multiplication between the contrast vector and a column vector of parameter estimates (i.e., $H\beta$) yielded a scalar value representing the degree of co-linearity between H and β. Note that this matrix multiplication is equivalent to taking the dot product between H and β which returns the magnitude of the projection of β onto H.

Hypotheses 2 and 4, involved testing differences between clustered and non-clustered conditions averaged over the 7 retention intervals. Accordingly, contrast vectors for these hypotheses weighted parameter estimates by their relative contributions to the clustered vs non-clustered effect. In both hypotheses 2 and 4, one fixed effect parameter contributed to the effect of clustering across all retention intervals and so was weighted with a factor of 7. Six other parameters each contributed to one of the delayed retention conditions and so was weighted by a factor of 1. Given these weightings, the contrast vector was then scaled to have a unit length by dividing each element by the overall magnitude.

Bayesian inference

In testing our *a priori* hypotheses, we computed BF_{10} as follows:

$$BF_{10} = \frac{\int_{\theta\in\Theta} \Pr(Data|H_1,\theta)\cdot\pi_1(\theta)\ d\theta}{\Pr(Data|H_0)} \qquad \textit{Eq. S9}$$

$\Pr(Data|H_1,\theta)$ is a normal distribution encoding the likelihood of the model parameters in θ under the alternative hypothesis (H_1), and Θ denotes the set of all possible parameters for H_1 (i.e., the parameter space). Additionally, π_1 refers to the prior distribution of these parameters. We used a Cauchy distribution as the prior π_1, see[5]:

$$\pi_1(\theta) = \frac{\Gamma(\frac{1+d}{2})\cdot\gamma}{\Gamma\left(\frac{1}{2}\right)\cdot\pi^{\frac{d}{2}}\cdot\left(\gamma^2+\sum_{i=1}^{d}{\theta_i}^2\right)^{\frac{1+d}{2}}} \qquad \textit{Eq. S10}$$

Where Γ denotes the gamma function, d is the dimensionality of the Cauchy distribution (i.e., the model degrees of freedom which is 1 for all *a priori* hypotheses), and γ is the Cauchy scale parameter. Note that π on the right-hand side of *Eq. S10* refers to the circle constant. Across each of our hypotheses, we fixed $\gamma = 0.555$.

In order to evaluate $\Pr(Data)$ in both the denominator and numerator of *Eq. S9*, the parameters returned by each GLMM (β) were multiplied by the contrast vector under test (H, i.e., the vectors listed in tables 1 and 2). This resulted in raw effect sizes ($z = H\beta$) that were standardised in order to be consistent with our Cauchy prior. This was achieved by dividing out the standard deviation of z obtained by multiplying the population covariance matrix (denoted $\boldsymbol{C}$) with H, and then taking the square root: $\sqrt{(H\boldsymbol{C}H^T)}$, where T represents the transpose operator. Finally, the variance for the normal distribution that encodes $\Pr(Data)$ was given by scaling the variance of the sampling distribution by the same standard deviations used previously. Using these statistics, both $\Pr(Data|H_1, \theta)$ and $\Pr(Data|H_0)$ were evaluated with the latter being the height of this distribution at the zero vector.

As well as providing Bayes factors, we report Cohen's *D* effect sizes for each hypothesis. This statistic was given by the following:

$$d = \sqrt{\frac{(H\beta)^2}{H\boldsymbol{C}H^T}} \qquad \textit{Eq. S11}$$

MATLAB functions implementing all the above computations are available at http://osf.io/8mzyc/.

Pilot study

We performed a lab-based, pilot study with 73 participants to validate our experimental design and generate estimated effect sizes for a sample size computation. This first involved parametrising the rate of forgetting for each measure of mnemonic information, in each condition. Subsequently, we used this parametrisation to simulate the main experiment and estimate the level of statistical power for a given number of participants.

The pilot study involved a similar task to that described above but did not include a subjective memory judgment at the end of each test trial. Also, instead of collecting data across 7 retention intervals, the pilot was limited to 3 retention intervals; one immediate test condition (0 hrs; *n* = 36), and two delayed test conditions - 24 hrs (*n* = 17) and 168 hrs (i.e., 7 days, *n* = 20). Given this data, we then performed the statistical analyses described previously with the exception that each mixed-effects model only included two delayed retention regressors. Supplementary Figure 1 displays mean estimates of I_t, I_p and I_k in each condition, and test statistics relating to each of our principal hypotheses are listed in Supplementary Table 1. These pilots' results provide evidence in favour of each of our a priori hypotheses (*BFs* > 6).

We also acquired online pilot data for the immediate test condition (0 hrs; *n* = 27), that showed comparable levels of performance and variability (in standard deviation units) relative to the lab-based pilot data: ***Clustered condition*** - *Online*: I_p= 0.493 (0.202), I_k= 1.187 (0.379); *In-lab*: I_p= 0.587 (0.392), I_k= 1.346 (0.455); ***Non-clustered condition*** - *Online*: I_p= 0.377 (0.166), I_k= 1.404 (0.566); *In-lab*: I_p= 0.499 (0.355), I_k= 1.404 (0.518).

Parametrisation of forgetting

Given the pilot data, we used a model of exponential decay to predict the rate of forgetting for I_t, I_p and I_k in the main experiment. Exponential decay is commonly used to model forgetting and is known to provide a good fit to behaviour in both short-term and long-term memory experiments[6]. Based on our mean estimates of I_p and I_k at each timepoint, we fitted the following model to these measures for clustered and non-clustered conditions (separately):

$$y(t) = \alpha + \beta \cdot \exp(-\lambda \cdot t) \qquad \textit{Eq. S12}$$

Where, t denotes the length of the retention interval (in hours), and y denotes the measure of mnemonic information being modelled (i.e., I_t, I_p or I_k in either the clustered or non-clustered condition). The free parameters α, β, and λ were estimated via the nonlinear least squares fitting method implemented in the MATLAB curve fitting toolbox. The fit of this model across each measure and condition was good; R^2 = .984.

Sample size computation

We ran simulations of the main experiment to estimate the sample size that would be required to achieve Bayes factors greater than 10 in favour of our a priori hypotheses. To do this, we used the above parametrisation of forgetting to generate mean estimates of I_t, I_p and I_k for the clustered and non-clustered conditions across all 7 retention intervals (Supplementary Figure 1). These means were then converted into hypothesised parameter estimates for the two GLMMs that constitute the main analysis. Variance-covariance matrices for these parameter estimates were also computed from the pilot analyses. Here, covariance components relating to each model term were pooled across retention intervals, and then redistributed into a larger matrix that included additional rows and columns for each of the 7 retention intervals. Finally, we rescaled these covariance matrices to reflect different samples sizes and performed Bayesian test for each of our five hypotheses. This revealed that a sample size of ~26 participants per retention interval condition should have yielded BF_{10} statistics greater than 10 (given the effect sizes we observed in the pilot study).

Kernel density estimation

We produced kernel density estimates characterising the distribution of location responses[7]. These estimates served three purposes: *(1)* to plot average distributions of angular errors in each condition (as in Figure 2B), *(2)* to plot the spatial distribution of responses relative to the experimentally imposed pattern in the clustered condition (as in Supplementary Figure 2), and *(3)* to compute the Kullback–Leibler divergence (D_{KL}) between the spatial distribution of participants' responses, and the pattern of studied locations in the clustered condition (see below). For a given set of n responses ($\widehat{\theta_\iota} \in \widehat{\Theta}$; e.g. all responses to clustered trails from one participant), the kernel density estimates (f_{kd}) at each position (t) in the interval $(-\pi, \pi]$ is given by the following:

$$f_{kd}(t \mid \widehat{\Theta}, k) = \frac{1}{n} \cdot \sum_{i=1}^{n} f_{vm}\left(\left(t - \widehat{\theta_\iota}\right) \bmod 2\pi \,\middle|\, k\right) \qquad \textit{Eq. S13}$$

As before, f_{vm} denotes the probability density function for a von Mises distribution with a mean parameter of 0 and concentration of k. Here, k acts as a smoothing parameter, often referred to as the bandwidth, that spreads the density function around each response in θ. For all uses in the current study, k was set to 2 as this provided smooth and reliable estimates in general (although, we note that the choice of k did not significantly alter the results). Importantly, depending on the purpose of the kernel density estimates, the responses in $\widehat{\Theta}$ were either angular errors in each condition, or angular differences between responses and the mean position of the experimentally imposed spatial cluster. The former case allowed us to estimate kernel density functions of angular errors. The latter allowed us to estimate spatial density functions of the responses themselves with location $t = 0$ corresponding to the centre of the cluster.

Kullback–Leibler divergence

Once the spatial distribution of responses had been estimated ($f_{kd}(t \mid \widehat{\Theta}, 2)$, see above), the Kullback–Leibler divergence[8] (D_{KL}) between this and the experimentally imposed pattern in the clustered condition was given by the following:

$$D_{KL}(\widehat{\Theta}) = \int_{-\pi}^{\pi} f_{vm}(t|2) \cdot \log\left(\frac{f_{vm}(t|2)}{f_{kd}(t \mid \widehat{\Theta}, 2)}\right) dt \qquad \textit{Eq. S14}$$

Note that the value of 2 used as a parameter for the kernel density estimate (f_{kd}) donates the bandwidth of the kernel. In contrast, the value of 2 used as a parameter for the von Mises probability density function (f_{vm}), reflects the concentration of the experimentally imposed spatial pattern.

Supplementary References

1. Fisher, N. I. *Statistical Analysis of Circular Data*. *Book* (1993). doi:10.1017/CBO9780511564345.

2. Best, D. J. & Fisher, N. I. The bias of the maximum likelihood estimators of the von mises-fisher concentration parameters. *Commun. Stat. - Simul. Comput.* **10**, 493–502 (1981).

3. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).

4. Kass, R. E. & Raftery, A. E. Bayes Factors Bayes Factors. *J. Am. Stat. Assoc. ISSN* **90**, 773–795 (1995).

5. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56**, 356–374 (2012).

6. Rubin, D. C. & Wenzel, A. E. One hundred years of forgetting: A quantitative description of retention. *Psychol. Rev.* **103**, 734–760 (1996).

7. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076 (1962).

8. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).