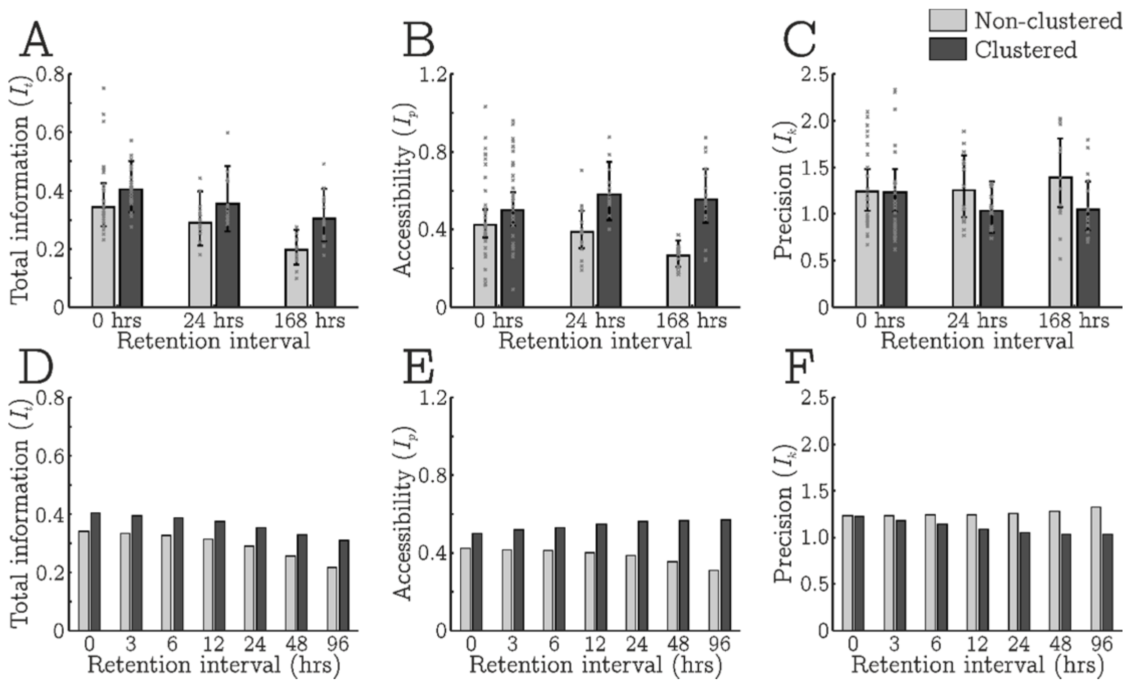
Supplementary Information

For “**Dissociating loss of memory accessibility and precision**”

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 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 Methods

Mixture model estimation

We first compute 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 starts by first assigning arbitrary random values to the parameters being estimated. The algorithm then progresses in two steps (an E-step and an M-step) that are repeated in sequence across multiple iterations. During the E-step, we compute a set of weightings (w_i) representing the probability that individual responses were based on memory retrieval (von Mises distributed errors). These weightings are 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 compute new values for each model parameter (the M-step). The parameter p is computed as follows:

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

To re-estimate the parameter k, we first compute 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).

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

Where j denotes the imaginary unit. The statistic r is 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 apply 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 repeat until the negative log-likelihood (NLL) of the model (i.e., the goodness-of-fit), converges 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 will be 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 will then be selected as the final model.

Assessing model fit

In cases where a participants retrieval probability is low ($p \lesssim$ 0.2), the EM algorithm may fail to converge or may incorrectly fit a wide von Mises distribution that is indistinguishable from a uniform (k ≈ 0.1). This latter case results in overinflated estimates of retrieval probability since the similarly shaped uniform and von Mises distributions will provide equal weightings to all data points (i.e., w ≈ 0.5 in all cases). 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 use 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 is computed as follows:

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

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 will take ΔBIC values of -10 or below to indicate that the model has properly converged and the parameters are 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 returns a ΔBIC greater than the -10 threshold, or fails to converge altogether, we attempt to identify a valid fit via an alternative search procedure. At first, this involves 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 will accept mixture model estimates identified in this way as long as the corresponding ΔBIC statistic is 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 will therefore exclude the data from participants when this is the case. If no mixture model can be fit to a participant's data such that the ΔBIC statistic is less than -10, the participant will be excluded from further statistical analyses.

Linear contrasts

Hypotheses 1, 3, and 5, involve 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 specify a 1-by-6 contrast vector, $H = [h_1, h_2, h_3, h_4, h_5, h_6]$, that evaluates differences between pairs of parameter estimates, and weights these differences by the time between retention intervals. Each element of H is 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 $7/6$ ensures that each delayed retention interval (i) is compared to the immediate retrieval condition (represented by the intercept term) as well as every other delayed condition. The resulting vector is then scaled to have a unit length by dividing each element by the overall magnitude. This results in a set of contrast weights that linearly decrease as a function of time. Consequently, performing a matrix multiplication between the contrast vector and a column vector of parameter estimates (i.e., $H\beta$) yields 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, involve testing for differences between clustered and non-clustered conditions averaged over the 7 retention intervals. Accordingly, contrast vectors for these hypotheses should weight parameter estimates by their relative contributions to the clustered vs non-clustered effect. In both hypotheses 2 and 4, one fixed effect parameter contributes to the effect of clustering across all retention intervals and so is weighted with a factor of 7. Six other parameters each contribute to one of the delayed retention conditions and so are weighted by a factor of 1. Given these weightings, the contrast vector is then scaled to have a unit length by dividing each element by the overall magnitude.

Bayesian inference

In testing our *a priori* hypotheses, we compute 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 will use 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 will fix $\gamma = 0.555$.

In order to evaluate $\Pr(Data)$ in both the denominator and numerator of *Eq. S9*, the parameters returned by each GLMM (β) needed to be multiplied by the contrast vector under test (H, i.e., the vectors listed in tables 1 and 2). This results in a raw effect size ($z = H\beta$) that must be standardised in order to be consistent with our Cauchy prior. This is achieved by dividing out the standard deviation of z that is 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 $\Pr(Data)$ is given by scaling the variance of the sampling distribution by the same standard deviations used previously. Given these statistics both $\Pr(Data|H_1, \theta)$ and $\Pr(Data|H_0)$ can be evaluated with the latter being the height of this distribution at the zero vector.

As well as providing Bayes factors, we will report Cohen's *D* effect sizes for each hypothesis. This statistic is 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/.](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 > 5).

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 for the main proposed 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 will 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 yield BF_{10} statistics greater than 10.

Supplementary references

1. Fisher, N. I. *Statistical Analysis of Circular Data*. *Book* (1993).

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).