nature human behaviour

Supplementary information

https://doi.org/10.1038/s41562-025-02390-4

Neural activations and representations during episodic versus semantic memory retrieval

In the format provided by the authors and unedited

Supplementary Analysis 1: Sample Size Determination

The sample size for the study was determined using a Bayesian "sequential design with maximal n"[26], where the maximal feasible number of participants, given time and funding, was n=100. Using this approach, the experiment was due to run in batches (with 40 participants in the first batch, and 10 participants in each additional batch), calculating the Bayes factor (BF) in favour of our main hypotheses (see Design table and "Data Analysis" section below) after each batch. For the purpose of the simulations, we used a maximal n=2,000 to fully appreciate the power of our design using various sample sizes. Once the simulations ran with the full sample size (of n=2,000), the results were capped and interpreted for smaller sample sizes (in this case, n=100).

Figure 1 shows the results of 10,000 iterations that used to determine the maximal number of participants that falls within our feasibility limits, but that would also be informative. Namely, for each iteration we generated two values for each sample consisting of n=40 participants, where the differences between these values are drawn from two normal distributions whose mean differed according to the expected effect size (see below). Note that the analysis of the data in our study was based on single trials, and so in reality, each participant had a lot more than two values. This renders the current estimation likely to be conservative. However, because the allocation of trials into bins (i.e., success / failure) can only be done after data are collected, and therefore the trial distribution was unknown when the simulations were conducted, this conservative approach was preferred. The simulated effect size was chosen based on a previous study[27], which used a design most similar to the one used here. In that study, the authors used a cued-recall paradigm to identify recall-related activity (via a univariate analysis, though effects sizes for this analysis are not reported) as well as event-specific activity (via a similarity analysis). For the similarity analysis, with n=20, the authors reported an interaction effect between the region of interest (ROI) factor (10 ROIs in their case) and the "match" factor (equivalent to our "trial type" factor), which translates to a medium-large effect size (Cohen's d = 0.67). Nevertheless, to accommodate possible over-inflation of the effect size (e.g., owing to publication bias), we set d=0.5 as a lower, more conservative value for our simulations.

We then z-scored the outcome (y) and fitted a BRMS model to the generated data:

$$y \sim x + (1 \mid subject)$$

where x is a dummy-coded input variable for the two levels. We fitted the BRMS models using 36,000 samples (of which the first 3000 were warm-up samples), with a unit normal prior, N(0, 1), for the

mean effect of x (and intercept) and a Student's T(3, 0, 10) prior for the standard deviation of x. BFs were approximated with the Savage-Dickey ratio[28], calculated by extracting the posterior samples for x, and using the polspline R package[29] for logspline density estimation of the posterior point density at zero. We then used the base function dnorm() to do the same for the prior.

The stopping criteria for the iteration was set to BF10 > 10 or < 1/10. Thus, if the BF was greater than 10 (which would indicate support for our main hypothesis that the effect of x ≠ 0, also called the "alternative hypothesis", BF10) or lower than 1/10 (which would indicate support for the null hypothesis that effect of x = 0, BF01), the iteration ceased, and the next one was initiated. If, however, the BF10 was < 10 and > 1/10, the iteration repeated with another batch of 10 (n=50, n=60, n=70…), until a BF10 of 10 or 1/10 was obtained, or until the maximum number of participants had been reached.

Using this effect size of d=0.5, 97% of the iterations resulted in conclusive support for H1 (BF10 > 10). Only 0.02% erroneously resulted in conclusive support for the null (BF10 < 1/10), while the remaining 2.68% were inconclusive. Moreover, for 99.8% of the iterations, the lower bound of the 95% credibility interval (CI), i.e., lower 2.5% of the posterior distribution for the effect of x, was above zero. This result is comparable to the power for rejecting the null hypothesis in frequentist statistics. This simulation indicated that we have a high probability (>95%) of finding evidence for our hypothesis (that is, that episodic and semantic memories are dissociable) if the effect size is at least 0.5.

We further calculated the probability that our study would support the null hypothesis, if d=0 were true (i.e., no difference between episodic and semantic memories). We ran an additional simulation with data generated with d=0 and a maximum n=2,000. This simulation showed that to obtain conclusive results for the null at a probability > 95%, we would need a sample size of n=440 which is not practical for the current study. Nevertheless, at n=100, 50% of the iterations produced conclusive support for the null (BF10 < 1/10; 76% for BF10 < 1/6), and only 0.0091% of the iterations erroneously resulted in conclusive support for H1 (BF10 > 10). Furthermore, the results indicated that at n=100, for 94.9% of the iterations, the 2.5% and the 97.5% bound of the 95% CI straddled zero. Thus, the simulations indicated that in the (unpredicted) case of no true effect, we were still likely to get evidence in favor of the null. In any case, throughout the study, BFs that do not reach 1/10 (or 10) were not interpreted as conclusive evidence for either hypothesis.

Given the above results, we set the stopping criteria to be n=100, or BF10 > 10, or BF01 < 1/10 for either of the two main analyses (univariate or similarity) for the a-priori-defined networks of interest.

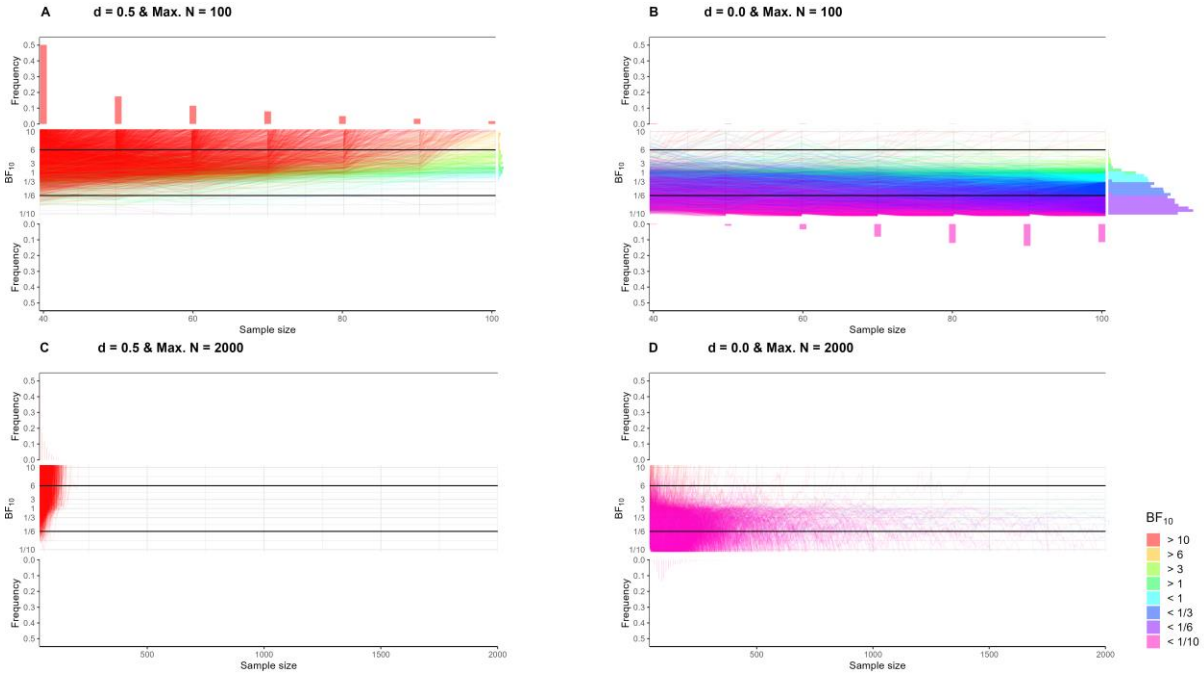


Figure SA1.1. Results of simulations for sample-size determination. Two effect sizes were simulated: 0.5 (medium effect size; left, Panels A & C) and 0 (null effect; right, Panels B & D). Here we show the results of the simulations when Max N was set to 100 participants (top, Panels A & B) and 2000 (bottom, Panels C & D). For each simulation, each line in the middle panel indicates one iteration (out of 10,000), with line color indicating the BF that was obtained (see legend on the right). Top and bottom panels indicate the proportion of iterations that were ceased for each sample size with BF>10 (top) or BF<1/10 (bottom). For example, for d=0.5, ~25% of the iterations were terminated at n=40 because a BF>10 was obtained. Additional ~18% were terminated at n=50, and so on). The right panel of each simulation shows the distribution of BFs obtained for inconclusive iterations.

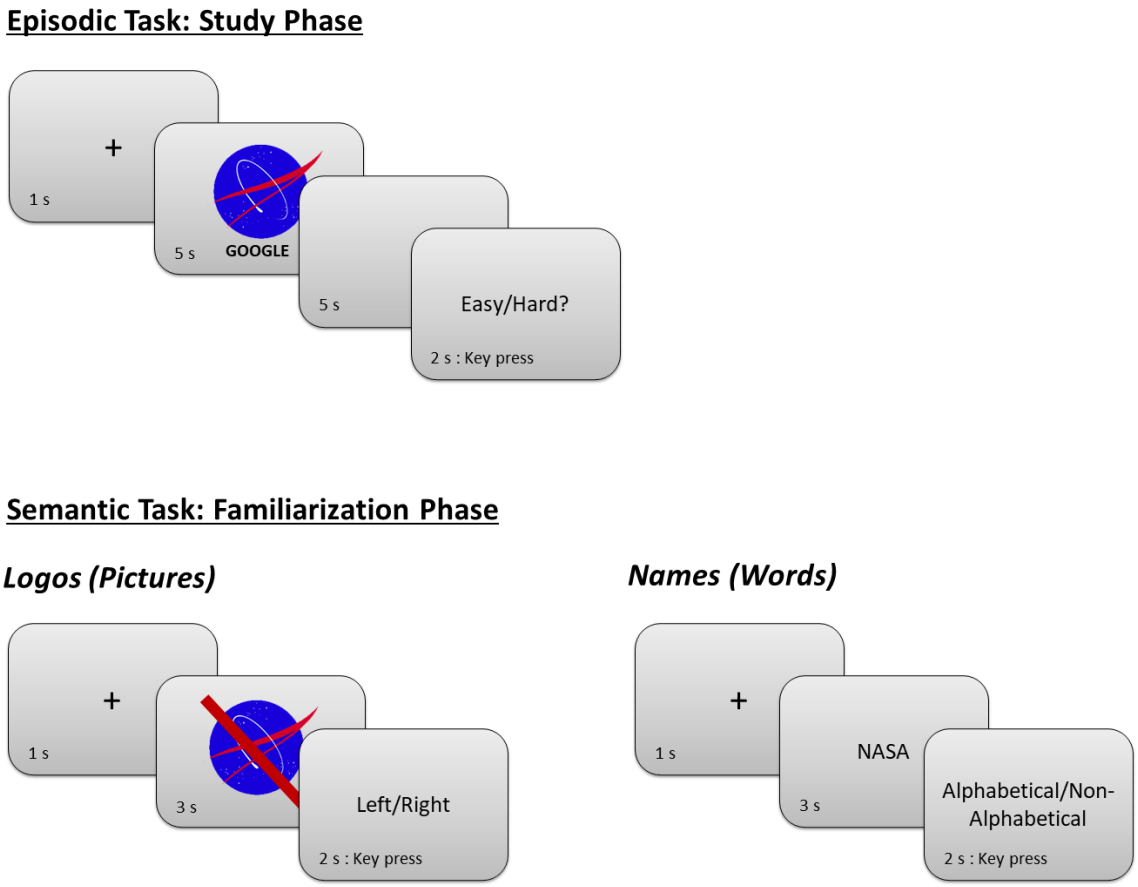
Supplementary Analysis 2: Validation of ICA-denoise

We used the "Episodic Network" and the "Semantic Network" used in our main analyses as our regions of interest, and with a custom MATALB script built on SPM tools ('roi_extract.m' from https://github.com/MRC-CBU/riksneurotools/blob/master/Util/), extracted ROI information. For each ROI, we extracted two metrics: 1) activation magnitude, i.e., the difference in GLM parameter estimates ("betas") for task (stimuli presentation) against rest, and 2) activation precision, i.e., the t-statistic for that difference, i.e., the activation magnitude scaled by an estimate of the uncertainty of that difference, which also depends on the noise in the data. Then, to determine the reliability of these two metrics, we performed t-tests across participants to compare combined and combined + ICA-denoised data. This was done separately for the episodic and semantic task. The results in the table below show that the combined + ICA-denoise procedure was consistently advantageous compared to the combined procedures, in all tasks, ROIs and metrices.

Task	**Episodic**		**Semantic**	
Network	**Episodic**	**Semantic**	**Episodic**	**Semantic**
	Magnitude			
T-Stat	5.4***	5.9***	7.1***	8.24***
	Precision			
T-Stat	11.42***	10.93***	8.75***	7.8***

Table SA2.1. T-test statistical comparisons between combined and combined + ICA-denoised data using semantic and episodic networks as regions of interest (ROIs), in the semantic and episodic task, for activation magnitude (contrast of betas) and activation precision (statistical T-values). dfs=35; ***p<0.001 Bonferroni-corrected for 8 tests).

Supplementary Figure 1: Schematic Illustration of Study Trials



Supplementary Figure 1. Schematic illustration of the unscanned study phase in the episodic task (top) and the familiarization phase in the semantic task (bottom). During the study phase of the episodic task, a picture of a logo was presented together with an unrelated brand name. Participants were asked to come up with a story linking the two of them together, and to indicate whether it was easy or hard for them to come up with the story. During the familiarization phase of the semantic task, participants viewed logos bisected by a red line and brand names in separate blocks. For logos, they were asked to indicate whether a bisecting line appears more to the left or the right of the picture. For names they were asked to indicate whether the first and last letters are in alphabetical order.