nature portfolio

Corresponding author(s): *Double-anonymous peer review submissions: write DAPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	data were collected using E-Prime (Version 2.0), Psychology Software Tools, Inc. (2016). https://www.pstnet.com
Data analysis	data were preprocessed / analyzed using FSL (v5.0.11), AFNI (v18.3.03), Tenada toolbox (v0.x), SPM12, R v4.2.0 with toolboxes BRMS v2.17.0, ggplot2 v3.4.1, tidybayes version 3.0.2, BayesFactor v0.9.12.4.3, tidyverse v2.0.0, plyr v1.8.7, and custom Matlab and R code fully available on OSF (https://osf.io/dm47y/)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw neuroimaging data, behavioral data, and lab logs are available on https://openneuro.org/datasets/ds004495/. Processed data, pilot data, and materials are available on https://osf.io/dm47y/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Our final sample (N=36) included female (25) and male (11) adults (mean age = 24.1, STD = 5.12, range 18-35). Participant's gender was determined based on self-reporting (options included "male", "female", "other"). As preregistered, participants were pseudo-randomly assigned into the various counterbalancing conditions, in order to keep the mean age and gender distribution across the various experimental lists roughly equated.
Reporting on race, ethnicity, or other socially relevant groupings	We have not collected this information
Population characteristics	Our final sample (N=36) included female (25) and male (11) adults (mean age = 24.1, STD = 5.12, range 18-35), who were native English speakers, MR-compatible, with normal or corrected-to-normal vision, and not diagnosed with attention deficit hyperactivity disorder, dyslexia, or any other developmental or learning disabilities.
Recruitment	Forty participants were recruited from the MRC Cognition and Brain Sciences' SONA system, or from word-to-mouth. This initial batch was sufficient for the BF for the null to exceed our stopping criterion (i.e., BF10 < 1/10) for both the univariate and similarity analyses (see Results). Data exclusion followed the criteria in our Stage 1 Report, which specifies that participants who are excluded prior to the analysis of their data will be replaced with others. According to the report, reasons for pre-analysis exclusion were failure to complete the task, failure to arrive to the second session, noncompliance with inclusion criteria, faulty equipment, and experimenter error. Accordingly, 5 participants were excluded prior to the analysis of their data; one due to an incidental MRI finding, two due to technical issues, one due to experimenter error in loading the sessions, and one who did not show up for the second session, and were replaced with others.
Ethics oversight	This study was approved by a local ethics committee (Cambridge Psychological Research Ethics Committee reference PRE.2020.018).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	quantitative study
Research sample	Participants were recruited through the MRC Cognition and Brain Sciences Unit's SONA system and via word-of-mouth. While the sample is not fully representative of the general population, SONA enables community-based recruitment, meaning that participants come from a variety of backgrounds and professions—not just university students. We recruited people aged 18-35, who were native English speakers, MR-compatible, with normal or corrected-to-normal vision, and not diagnosed with attention deficit hyperactivity disorder, dyslexia, or any other developmental or learning disabilities.
Sampling strategy	Participants were recruited through the MRC Cognition and Brain Sciences Unit's SONA system and via word-of-mouth. The sample size for the study was determined using a Bayesian "sequential design with maximal n", 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 favor of our main hypotheses. Following simulations, the stopping criteria was set 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.
Data collection	The experimental protocol consisted of two sessions, each comprising two phases. The first phase took place in a computer lab, where participants completed a computerised task in the presence of the researcher only. The second phase involved MRI data acquisition using a Siemens 3T PRISMA scanner with a 32-channel head coil, conducted in a shielded MRI suite. Both the researcher and a radiographer were present during the scanning session. The researcher was not blinded to the experimental conditions or the study's hypotheses.
Timing	17 Aug 2021 - 09 Dec 2021
Data exclusions	Data exclusion followed the criteria in our Stage 1 Report, which specifies that participants who are excluded prior to the analysis of their data will be replaced with others. According to the report, reasons for pre-analysis exclusion were failure to complete the task, failure to arrive to the second session, noncompliance with inclusion criteria, faulty equipment, and experimenter error. Accordingly,

5 participants were excluded prior to the analysis of their data; one due to an incidental MRI finding, two due to technical issues, one due to experimenter error in loading the sessions, and one who did not show up for the second session, and were replaced with others. Our Stage 1 Report further indicates that participants who were excluded following the analysis of their data will be removed from the relevant analyses but will not be replaced. Potential reasons for post-analysis exclusion were lack of trials in one or more experimental conditions (due to floor/ceiling performance), increased movements (2 SDs or more than the mean motion across all participants), and/or motion that resulted in reduced coverage of key locations (within a-priori-defined networks, as detailed below). Accordingly, four participants were excluded following the analysis of their data but were not replaced. These included one exclusion due to lack of trials in one of the experimental conditions, one due to increased head movements, and two exclusion of two participants for which the "tedana" denoising algorithm (detailed below) failed to converge during preprocessing.

Non-participation

No participants declined participation. One participant did not show up for the second session, and was replaced by another.

Randomization

Participants were not allocated into experimental groups (fully within-subject design). Participants were pseudo-randomly assigned into the various counterbalancing conditions, in order to keep the mean age and gender distribution across the various experimental lists roughly equated.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Plants

Seed stocks

N/A

Novel plant genotypes

N/A

Authentication

N/A

Magnetic resonance imaging

Experimental design

Design type

event related

Design specifications

The design included 3 scanned tasks: Episodic retrieval, Semantic retrieval, and Control. The Episodic retrieval task included logo-cue trials and name-cue trials. Each logo-cue trial started with a jittered fixation cross (1-8 s; mean=4), followed by the presentation of a picture for 1 s, a 5 s blank screen, a response screen shown for 3 s (verbal response), and a second response screen shown for 2 s (key press). After completing sub-block of 48 logo-cue trials, a sub-block of 48 name-cue trials ensued, corresponding to the same stimuli. In each name-cue trial, after a jittered fixation (1-2 s; mean=1.5), a word was presented for 1 s, followed by a 5 s blank screen, and 2 s response screen (key press). The test phase was comprised of 3 blocks (runs). In each block, a sub-block of 48 logo-test trials was displayed, followed by a sub-block of 48 brand-test trials. The structure of the Semantic retrieval task was identical, but included 2 (instead of 3) blocks. For the control task, each trial started with a jittered fixation cross (1-8 s), followed by a letters-display presented for 1 s, a 5 s blank screen, a 3 s response screen (verbal) and a second 2 s response screen (key press). The control task included two blocks (runs) of 40 trials. The Episodic task took place in one day, and the Semantic and Control task took place together on another day.

Behavioral performance measures

Following behavioral pilots, we expected participants to reach ~50% accuracy in each task. Nevertheless, we took

Behavioral performance measures

precautions to ensure that performance remained interpretable and is not at floor or at ceiling when participants performed the task in the scanner. In our Stage 1 Report, we indicated that performance will be assessed after data are collected from the first 5 participants, and the task(s) will be revised if floor/ceiling effects are detected. Accordingly, performance was assessed after data were collected from the first 5 participants (one for each of the 5 stimulus lists). Floor performance was defined as when the average of successful trials was < 10%, and ceiling performance were defined as when the mean of successful trials was > 90% in any of the tasks. Assessment of the data indicated that there were no floor/ceiling effects in any of the tasks: In the episodic task, mean accuracy across the first 5 participants was 36% (STD = 16%, range = 15%-58%), in the semantic task, mean accuracy was 45% (STD = 15%, range = 24%-68%), and in the control task, mean accuracy was 41% (STD = 0.044%, range = 35%-46%).
Statistical analyses were performed in R and RStudio. Data were analyzed using Bayesian mixed-models analyses that accommodate both within- and between-subject variability. This approach is particularly recommended for unbalanced data (an unequal number of trials in each condition, which we have here due to the post hoc division of trials into successful and failure trials. Behavioural data were analysed using a logistic regression model with binomial family link function. Trial outcomes (1 for success, 0 for failure) were submitted to a Bayesian logistic regression model which included task (episodic, semantic, control) as a fixed factor, and a subject-specific intercept and slope for this factor as the random part of the model. We fitted the model with the following formula, where "(... |x)" indicates random effects: outcome ~ task + (1 + task | subject). We fitted the model across 36,000 samples of which the first 3000 were warm-up samples, using the Bayesian Regression Models (BRMS) R package.

Acquisition

Imaging type(s)

Structural, Functional

Field strength

3T

Sequence & imaging parameters

The same acquisition and preprocessing protocol were used for all tasks. MRI data were collected using a Siemens 3 T PRISMA system with a 32-channel head-coil. Structural images were acquired with a T1-weighted 3D Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence [repetition time (TR) = 2250 ms; echo time (TE) = 3.02 ms; inversion time (TI) = 900 ms; 230 Hz per pixel; flip angle = 9°; field of view (FOV) 256 × 256 × 192 mm; GRAPPA acceleration factor 2]. Functional images were acquired using an echoplanar imaging (EPI) sequence with multi-echo (4) multi-band (2; MEMB) acquisition. We used a multi-echo (ME) sequence in order to recover signal from the ATL, which is a region associated with semantic memory, but suffers from susceptibility artifacts in gradient-echo fMRI32–34; we used multi-band (MB) to compensate for the longer TR required for multi-echo acquisition. For the episodic task, volumes were acquired over 3 runs. For the semantic and the control tasks, volumes were acquired over 2 runs. Each volume contained 46 slices acquired in interleaved order within each excitation band, with a slice thickness of 3 mm and no interslice gap (TR = 1792 ms; TEs = 13 ms, 25.85 ms, 38.7 ms, and 51.55 ms; flip angle = 75°; FOV = 192 mm × 192 mm; voxel-size = 3 mm × 3 mm × 3 mm). Field maps for EPI distortion correction were also collected (TR = 541 ms; TE = 4.92 ms; flip angle = 60°; FOV = 192 × 192 mm).

Area of acquisition

Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software

All raw DICOM data were converted to nifti format using dcm2niix. The T1 data were processed using FSL (v5.0.11) and subjected to the 'fsl_anat' function. This tool provides a general processing pipeline for anatomical images and involves the following steps (in order): 1) reorient images to standard space ('fslreorient2std'), 2) automatically crop image ('robustfov'), 3) bias-field correction ('fast'), 4) registration to MNI space ('flirt' then 'fnirt'), 5) brain extraction (using fnirt warps), and 6) tissue-type segmentation ('fast'). Images warped to MNI space were visually inspected for accuracy.
The functional MEMB data were pre-processed using a combination of tools in FSL, AFNI (v18.3.03) and a python package to perform TE-dependent analysis. Images were de-spiked (3dDespike), slice-time corrected (3dTshift, to the middle slice), and realigned (3dvolreg) with AFNI.

Normalization

Denoised and combined images were then averaged, and the mean coregistered to T1 (flirt) and warped to MNI space using fnirt warps and flirt transform. For whole-brain activation analyses, smoothing was applied using an 8 mm FHWM Gaussian kernel.

Normalization template

MNI152

Noise and artifact removal

Following preprocessing with FSL and AFNI, images were submitted to the "tedana" toolbox (max iterations = 100, max restarts = 10). Tenada takes the time series from all the collected TEs, decomposes the resulting data into components that can be classified as BOLD or non-BOLD based on their TE-dependence, and then combines the echoes for each component, weighted by the estimated T2* in each voxel, and projects the noise components from the data. Because the benefits of the ICA-denoising procedure were not yet established for task data when our Stage 1 report was accepted (although they were established in a recent paper, specifically designed to address this issue), we stated then that we will analyze the data using two methods—combined and combined + ICA-denoise—and formally compare the results. This analysis showed that the combined + ICA-denoise procedure was consistently advantageous compared to the combined procedures.

Volume censoring

We did not apply volume censoring. However, we included movement parameters to capture movement-related artifacts, and excluded a participant with increased head movements (as preregistered)

Statistical modeling & inference

Model type and settings

We employed a univariate analysis (detailed here) and a similarity analysis (detailed below) The univariate analysis was conducted for logo-cue trials classified as "success" and "failure" trials. It was conducted separately for a-priori-defined networks and data-defined clusters. SPM12 in Matlab was used to construct GLMs for each participant separately for each run and for each task. These first-level GLMs included two separate regressors for each response type of interest for logo-cue trials (success, failure), a regressor for excluded responses (filler trials, false alarms, and unnameable targets, see above), and a regressor for name-cue trials which are of no interest for this analysis. Predicted responses in these regressors were locked to the onset of stimulus presentation. The GLM further included a regressor for the verbal response and a regressor for the motor response to the slide (locked to the motor response. Each of these regressors were generated with a delta function convolved with a canonical hemodynamic response function (HRF). Six subject-specific movement parameters were also included to capture residual movement-related artifacts. The GLM was fitted to the data in each voxel. The autocorrelation of the error was estimated using an AR(1)-plus-white-noise model, together with a set of cosines to high-pass the model and data to 1/128 Hz to remove low-frequency noise, fitted using restricted maximum likelihood. The estimated error autocorrelation was then used to "prewhiten" the model and data, after which ordinary least squares were used to estimate the model parameters.
Group level analyses were conducted using mixed-effect models. Importantly, rather than averaged estimates across participant/condition, the mixed-models require estimation of the BOLD response for each trial. To get these estimations, we used the selected locations. For each participant, we extracted time series data from these locations using the first eigenvector across all voxels in each location, and adjusting for effects of no interest such as those captured by the motion regressors. To estimate the BOLD response to each trial from this time series, we used the least-squares separate (LSS-N) approach, where "N" is the number of conditions. LSS-N fits a separate general linear model (GLM) for each trial, with one regressor for the trial of interest and one regressor for all other trials of each of the N conditions. Only single trials of the conditions of interest were estimated this way. LSS implements a form of temporal smoothness regularization on the parameter estimation. The regressors were created by convolving a delta function at the onset of each stimulus with the HRF. The parameters for the regressor of interest were then estimated using ordinary least squares, and the whole process was repeated for each separate trial of interest, and for each location.
The resulting betas were submitted to a Bayesian linear mixed-model which included two locations (either predefined or data driven, depending on the analysis), the two critical tasks (episodic, semantic), response type (success, failure), and their interactions as fixed effects. The model further included subject-specific slopes for each factor and a subject-specific intercept as the random part of the model. The "ROI" factor always included two locations. We used a linear regression model with Gaussian family link function, and fitted the models across 36,000 samples of which the first 3000 were warm-up samples, with the following formula, with a unit normal prior set for the intercept and for the mean effect of x: betas ~ ROI * task * response type + (1 + ROI + task + response type | subject).

Effect(s) tested

We tested our prediction of greater recall success effect (success > failure) in the episodic task in the episodic network, but greater recall success effect in the semantic task in the semantic network for a-priori defined networks. We also tested a two-tailed version of the same interaction between ROI and task for the data-defined clusters. The present of any such interaction involving two locations would support our hypothesis that the episodic and semantic systems are dissociable.

Specify type of analysis: ☐ Whole brain ☒ ROI-based ☐ Both

Anatomical location(s)

We used two definitions to identify locations of interest. The first was an a-priori definition of networks, based on functional data from previous studies. For this definition, we identified two networks of interest: a core-recollection network that is applicable in various episodic tasks, regardless the nature of the recollected content, and a semantic network, comprised of regions that show consistent activation in tasks that require semantic processing, combined with the ATL from a recent fMRI study using a paradigm with enhanced ATL signal coverage. In order to maximize our ability to detect potential differences, we excluded any overlapping voxels that are shared between the two networks.
The second definition was based on data from the current study (i.e., data-defined clusters). For this definition, we ran a whole-brain analysis and identified locations where the recall success effect (success-failure) in the two critical tasks (episodic and semantic) was greater than the success effect in the control task ($p < 0.05$ FWE cluster-level corrected, with voxel-level threshold at $p < 0.001$ uncorrected). Note that this contrast is orthogonal to the difference between episodic and semantic success effects, so does not bias the subsequent comparisons of these two. The sample used for this analysis consisted of 32 participants, following the exclusion of 4 additional participants: 2 for which behavioral data for the control task were not available due to problems with the recordings, and 2 for which MRI data for the control task were not recorded due to technical issues. 2 clusters were obtained with this definition.

Statistic type for inference

(See Eklund et al. 2016)

We used trial time-series data extracted with LSS (details above). For the univariate analysis, trial time-series data were extracted from each ROI. For similarity analysis, trial time-series data were extracted from each voxel.

Correction

For the whole brain analysis used for data-defined clusters (see above) we used $p < 0.05$ FWE cluster-level corrected, with voxel-level threshold at $p < 0.001$ uncorrected. In an exploratory analysis with a revised threshold, we increased the statistical threshold to $p<.05$, corrected for family-wise error (FWE) using SPM's random field theory for peak-level inference.
For BRMS, in our Stage 1 report, we stated that to avoid an inflated false positive rate due to multiple comparisons, we will adjust the model's prior based on the number of comparisons that we perform. In practice, this was not required because our registered analysis only yielded two clusters, resulting in a single comparison. Nevertheless, in the exploratory analysis that used a revised threshold, we extracted data from each of 16 ROIs, and fitted the BRMS using different pairs of ROIs each time, until all pairs had been contrasted (120 pairs in total, one for each possible pair of regions). This exploratory analysis therefore included many comparisons, and so to avoid an inflated false positive rate, we adjusted the BF criterion for conclusive evidence. To this end, we used simulated data with a null effect to get the false positive rates for one comparison

and for 120 comparisons. Based on these results, we repeatedly adjusted the BF criterion until the false positive rate obtained with 120 comparisons was the most similar to that obtained with one comparison. Using this approach, we found that BFs should be adjusted by 0.0222, i.e., "corrected" BF10 = 10/0.0222 = 450.45.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

The similarity (multivariate) analysis was conducted for success trials in the two critical tasks (episodic, semantic), and was restricted to pairs of logo- and name-cue trials that corresponded to the same level of details (i.e., classified as "match+" or "match-" trials, see "Trial classification" section above). In addition, because pairs of logo-brand trials that correspond to the same event ("same trials") were always from the same block (run), each logo-cue trial was only correlated with an unassociated name-cue trial from the same block, to avoid confounding same/different with temporal lag between logo-cue and name-cue trials. 1st level GLMs included the same regressors as in the univariate analysis, with additional regressors for logo-cue and name-cue trials modelled as a 5 s boxcar function at the onset of the delay period. Similar to the univariate analysis, the similarity analysis was performed separately for each type of location definition. Nevertheless, instead of extracting a single time-course from each location, we extracted a time-course from each voxel within the locations. Moreover, the time of interest was the delay period where no perceptual information was presented (5 s following stimulus presentation, modelled as an epoch). For this analysis, we focused on the delay period rather than stimulus onset, as was done for the univariate analysis. Epoch data were extracted using the LSS-N approach described above. Following this extraction of single-trial time courses, successfully recalled logo-cue trials, and their corresponding name-cue trials, were considered to be trials-of-interest if labelled as 'match' trials (see "trial classification" section above). Other trials (mismatch, failure, false alarms or unnamable targets) were excluded.
Logo-brand similarity was then computed using Fisher-transformed correlation coefficients for (1) "same" pairs: success logo-cue trials classified as detailed / non-detailed and the name-cue trial which corresponds to the associated brand (i.e., from the study phase in the episodic task, or the actual associated brand in the semantic task), and (2) "different" pairs: success logo-cue trials classified as detailed / non-detailed and a different (unassociated) name-cue trial corresponding to the same level of detailedness, and presented at the smallest time-lag, either before or after, from the associated brand-test trial. These similarity coefficients were submitted to a linear mixed-model with ROI, task (episodic, semantic), trial type (same, different), and their interactions, as the fixed part of the model, and with subject-specific slopes (for each factor) and intercept as the random factors with the same priors described above: coefficient ~ ROI * task * trial type + (1 + ROI + task + trial type | subject)
We tested our prediction of greater similarity effect (same-different) in the episodic task in the episodic network, but greater similarity effect in the semantic task in the semantic network (for a-priori defined networks). We also tested a two-tailed version of the same interaction between ROI and task for the data-defined ROIs. We rationalized that such pattern would indicate that semantic and episodic content is represented in dissociable brain regions.