PNAS RESEARCH ARTICLE | NEUROSCIENCE
PSYCHOLOGICAL AND COGNITIVE SCIENCES

Identifying causal subsequent memory effects

David J. Halpern[a,1], Shannon Tubridy[a], Lila Davachi[b], and Todd M. Gureckis[a]

Edited by Richard Shiffrin, Indiana University Bloomington, Bloomington, IN; received November 8, 2021; accepted December 12, 2022

Over 40 y of accumulated research has detailed associations between neuroimaging signals measured during a memory encoding task and later memory performance, across a variety of brain regions, measurement tools, statistical approaches, and behavioral tasks. But the interpretation of these subsequent memory effects (SMEs) remains unclear: if the identified signals reflect cognitive and neural mechanisms of memory encoding, then the underlying neural activity must be causally related to future memory. However, almost all previous SME analyses do not control for potential confounders of this causal interpretation, such as serial position and item effects. We collect a large fMRI dataset and use an experimental design and analysis approach that allows us to statistically adjust for nearly all known exogenous confounding variables. We find that, using standard approaches without adjustment, we replicate several univariate and multivariate subsequent memory effects and are able to predict memory performance across people. However, we are unable to identify any signal that reliably predicts subsequent memory after adjusting for confounding variables, bringing into doubt the causal status of these effects. We apply the same approach to subjects' judgments of learning collected following an encoding period and show that these behavioral measures of mnemonic status do predict memory after adjustments, suggesting that it is possible to measure signals near the time of encoding that reflect causal mechanisms but that existing neuroimaging measures, at least in our data, may not have the precision and specificity to do so.

long-term memory | neuroimaging | causal inference | encoding | memorability

Significance

While decades of work has found robust neuroimaging correlates of successful encoding, it remains unclear whether these signals reflect causal mechanisms of memory. We develop a research design and analysis approach that allows statistical adjustment for a number of potential confounds that are well known to memory scientists but are typically not included in subsequent memory analyses. We find that several fMRI signals of subsequent memory proposed in the literature do correlate with memory performance three days later but do not after adjusting for confounders. However, we find that postencoding behavioral measures of mnemonic status do, suggesting that causal signals likely do exist. This work highlights the importance of adjusting for confounding factors when making causal claims from neuroimaging data.

What are the neural mechanisms that cause the encoding of lasting memories? The dominant method for identifying successful-encoding–related signals using noninvasive neuroimaging is the subsequent memory paradigm (1). This analysis approach compares neuroimaging signals, collected while a subject processes several items, based on later behavioral memory performance (e.g., comparing signals from items that were later remembered to those that were later forgotten). A signal that consistently differs between subsequently remembered and forgotten items suggests a link between the particular underlying activity and memory encoding. These differences, known as subsequent memory effects (SMEs), reliably appear in a number of brain regions, using a variety of memory tasks, imaging technologies and statistical techniques (1–3). Researchers have theoretically linked particular SMEs to specific latent cognitive or neural mechanisms of memory encoding such as attention, fatigue, representational fidelity, degree of associative binding, or match to personal schemas (4–14). In addition, SMEs have distinguished between multiple cognitive theories of memory encoding and learning (15, 16) and have been used practically to guide neural stimulation or optimization of learning for improving memory (17, 18).

While identifying signals that are associated with memory performance can be of interest itself, claims that the neural activity underlying SMEs reflect encoding mechanisms require that the activity be causally involved in encoding. For activity to be causal encoding activity, it means that, if we were able to manipulate the activity during the encoding of an event while holding all other external (i.e., nonneural) memory-related factors constant, memory performance would be better, on average, in one condition than the other.* However, identifying causal encoding activity experimentally is exceedingly difficult, if not impossible, because of our limited ability to precisely manipulate neural activity, particularly in humans where there are additional ethical constraints. Therefore, our best hope is to learn as much as we can about causal encoding activity from large-scale observational data, such as from neuroimaging studies. We argue that rather than avoid using causal language altogether to describe the results of observational studies, being upfront about the causal goal can allow us to evaluate to what extent our strategies for

*This definition does not hold neural activity constant because, for most signals, some neural activity is downstream and therefore would be a posttreatment variable which might confound causal identification (19, 20).

Author affiliations: [a]Department of Psychology, New York University, New York, NY 10003; and [b]Department of Psychology, Columbia University, New York, NY 10027

Author contributions: D.J.H., S.T., L.D., and T.M.G. designed research; S.T. performed research; D.J.H. contributed new reagents/analytic tools; D.J.H. and S.T. analyzed data; and D.J.H. and T.M.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: david.halpern@nyu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2120288120/-/DCSupplemental.

Published March 23, 2023.

identifying causal activity meet that standard (21–23). In some cases, a simple association, like an SME, can be interpreted causally, such as when there are no confounding variables that affect both the effect (memory performance) and the cause (neural activity). A key difficulty in the causal interpretation of SMEs is that several stimulus-, task- and context-related variables are known to do exactly that. For instance, the concreteness of a word is known to affect both a word's probability of recognition and recall on a list (24–26) as well as neural activity in several brain regions during visual presentation (27, 28). It is possible that some of this neural activity is involved only in word processing and does not directly affect the memory quality. But, in typical sets of word stimuli, signals reflecting this activity will vary with memory performance and thus be classified as a subsequent memory effect, regardless of their involvement in the encoding process.

Beyond concreteness, stimuli (including images, words, and videos) have also been shown to vary in their intrinsic memorability, i.e., the average performance for an item in a particular memory task in the population (29–34). The concern that confounding variables exist in subsequent memory analyses is not just theoretical: Memorability has been shown to drive a significant amount of neural activity in memory tasks during both encoding (35) and retrieval (36), including in regions typically also associated with SMEs. Therefore, analyses that rely on simply comparing signals between remembered and forgotten trials do not allow for clearly distinguishing between activity that is likely to be causal and activity that merely correlates with memory-predictive features of the stimulus or task.

Does this mean that progress toward uncovering the causal mechanisms for memory encoding from neuroimaging data is hopeless? We believe not. In the social sciences, causal claims are routinely made from observational data, using various approaches to measure and adjust for confounding variables (37–41).

The implicit causal model (*SI Appendix*) in past work on the SME assumes that a stimulus presentation is an exogenous event that initiates a causal chain leading to later memory behavior. This chain has intermediate steps that are neural, some of which are measurable (with some noise) using modern imaging techniques. However, we often do not measure all of these processes or at least not with significant precision, meaning that it is possible that none of the measured neural signals in any one study is part of the causal chain. Since it is difficult currently to directly manipulate the neural responses themselves, we can attempt to identify which of our neural signals are on the causal pathway using observational analyses. The standard SME demonstrates that some neural signals are correlated with memory behavior. If this neural activity is part of the causal chain, or otherwise influences variables that are on the chain, this causal model suggests that adding it as a predictor to all features of the initiating event that determine memory performance (such as the various item, task, and context variables) should increase the ability to predict memory (because it is closer in the chain to the final effect); if it is not part of the causal chain, then adding it will not increase prediction. This approach allows us to assess evidence for a causal role of the neural activity against the alternative hypothesis that the neural activity reflects stimulus-evoked activity that is simply correlated with memory performance on average. As emphasized by the causal model described here and in *SI Appendix*, some of the causal activity may be downstream from confounding variables. For instance, some neural activity that reflects concreteness may be on the causal pathway. However, to obtain evidence of this, we would need to observe that variation in this signal (and presumably the subject's subjective impression of a word's concreteness) predicts memory after adjusting for concreteness as measured in the population. If it does not, then the signal may just be a correlate of concreteness without being on the causal pathway to successful memory.

Two recent efforts, ref. 35, using fMRI with a visual recognition task, and ref. 42, using scalp EEG with a verbal free recall task, have attempted to investigate subsequent memory effects while statistically adjusting for item-level memorability and, in the case of ref. 42, effects of serial position, a task-level variable that is well known to affect probabilities of recall (43, 44). Both papers found that the adjusted subsequent memory effects appeared in a more limited set of brain regions (35) and had diminished predictive power (42). However, as mentioned above, there are many other known effects besides item memorability and serial position that affect the probability of successful encoding such as the distinctiveness or semantic similarity of the item relative to items studied nearby (45–48). Indeed, the probability of remembering an item in a particular serial position may depend on the item itself. Therefore, even some of the adjusted subsequent memory effects may in fact be confounded by stimulus, task, and context effects that drive both causal encoding activity and activity that is unrelated to encoding processes.

A major challenge, then, in identifying observational neuroimaging signals that plausibly reflect causal encoding activity is appropriately measuring all of the confounding variables. In every subsequent memory study we are aware of, item presentation is randomized to aid in generalizability. However, from the perspective of dealing with possible confounding factors, this presents a combinatorial problem. Estimating the total effect of all exogenous factors that affect memory, including stimulus variables, task variables (such as serial position), contextual effects of a stimulus' relationship to neighboring stimuli, and their possible interactions requires either strong assumptions about the functional form of these effects or an immense amount of data collection to estimate it nonparametrically.

In this study, we circumvent this combinatorial challenge by collecting a unique fMRI dataset of subjects performing a paired-associates verbal memory task where all subjects view the exact same items in the exact same order. This relatively unusual design, inspired by recent studies that measure brain responses across people to a single common experience (e.g., watching the same movie, 49, 50), allows us to precisely quantify the total effect of the experimenter controlled exogenous variables in our task.

Through a formal causal analysis (*SI Appendix*), we can see that this adjustment approach rules out activity that is associated with memory (an SME) but unlikely to be causally involved in encoding processes. Unfortunately, we cannot rule out that activity correlated with memory even after adjusting for confounding is not merely downstream from unmeasured activity that is truly causally involved in encoding (51). However, we argue that obtaining signals that measure activity downstream of causal activity is still valuable as they measure memory-predictive variables that may not be available from behavioral data alone. Therefore, this represents an advance over the unadjusted SME as identified signals provide a more solid foundation for advancing cognitive theory as well as applied goals such as building computer-aided systems to improve learning. In addition, it allows for using the relatively cheap observational data to more precisely target plausible sites for expensive stimulation studies. Following the psychometric and epidemiology literature

(52–56), we define a slightly weaker standard of an indicator of causal encoding activity or ICEA, that is, activity that is either on the causal pathway to encoding (i.e., causal encoding activity) or downstream from such activity.

Leveraging the above approach, the goal of the current study is to investigate whether several fMRI subsequent memory signals proposed in the literature appear to measure ICEA. While early fMRI studies on subsequent memory effects focused on demonstrating differences in univariate activation (**Activity**)[†] (1, 2), subsequent work expanded the types of signals studied to include multivariate patterns (3, 57). Deriving these signals involves comparing the spatial pattern of voxel activation across time points or across people commonly known as representational pattern similarity (58, 59), and the ability of these measures to predict memory in a variety of tasks has been extensively documented in the past decade (3). In general, we can categorize these signals into three classes: the pattern similarity of repeated presentations of the same item (Item Pattern Similarity or **IPS**) (15, 60–64), the pattern similarity of presentations of an item to other items studied in the same encoding period (Global Pattern Similarity or **GPS**) (35, 65–70), and the pattern similarity of the same presentation of an item to other participants in a study (Inter–Subject Pattern Correlation or **ISPC**) (49, 50, 71). To test these four signal types in a variety of brain regions while avoiding losing power due to large multiple comparisons corrections, we use a predictive modeling approach, testing whether many signals jointly predict recall performance in a regularized regression (72).

In sum, our approach to identifying fMRI features that measure ICEA combines a highly controlled memory task which allows for precise measurement of confounding variables with predictive modeling approaches that allow us to increase our power to estimate potentially small associations. Aside from presenting items in exactly the same order, our experiment has several other unique features. One particularly interesting application for causal encoding activity might be to track learning of novel information in an educational setting. It may be possible to use measurements of ICEA to adjust teaching plans in a way that will facilitate faster learning, as in automated tutoring systems, e.g., ref. 73. We therefore use a novel-language learning task in which participants learn associations between English words and their Lithuanian translations (74, 75). To allow for measuring within-item pattern similarity, our experiment (Fig. 1) consists of presenting 45 paired associates five times each during an initial encoding session. This increases overall memory

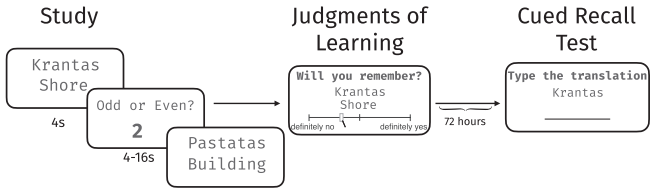

Fig. 1. In Phase 1, in the fMRI scanner, subjects first completed a paired associates learning task in which an English word and a Lithuanian word were simultaneously presented on the screen. In between trials, subjects completed an odd–even judgment about a series of numbers in order to prevent rehearsal. After the learning session, subjects made a judgment of learning about each of the paired associates outside of the scanner. In phase 2, 72 h after phase 1, subjects completed a cued recall test in which they were presented with a Lithuanian word and typed in the corresponding English word.

[†]Descriptions of components of models used in results figures are indicated in the main text using bold font.

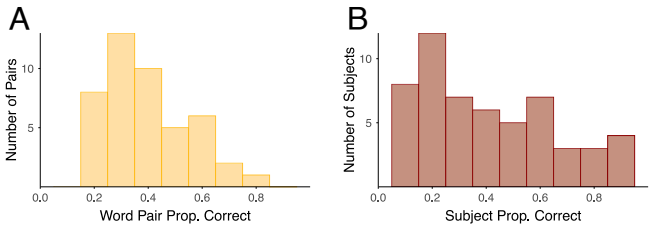

Fig. 2. (*A*) shows the distribution of word pair memorabilities (percent of subjects remembering the pair in a cued recall test) across all 45 word pairs. (*B*) shows the distribution of subject abilities over all 57 subjects.

accuracy and the robustness of the item-level measurements. As a comparison to fMRI measures of encoding, we additionally ask participants to provide subjective "judgments of learning" (JOLs), a common behavioral measure related to encoding confidence and metacognition (76–78), for each studied word pair. Finally, participants are tested for their cued recall performance 3 d after the initial encoding, ensuring that we are measuring signals of durable memory encoding.

We provide evidence that many previously identified SME signals are able to robustly predict memory formation in our task even in entirely held-out subjects. However, we also show that there is very weak statistical evidence that these signals reflect neural activity that we can claim is causally involved in memory encoding after accounting for possible confounds. Finally, we use the same framework to determine whether participants' own judgments of learning are related to ICEA. Repeating our analysis approach, we show that, perhaps surprisingly given the long delay between study and test, it is in fact possible to measure signals (e.g., subjects' behavioral responses) at encoding time that do reflect the quality of encoding processes. However, current fMRI indices of memory encoding do not yet have the required level of precision and specificity to measure them reliably.

Results

Behavioral Performance. We quantified both the average performance of each of 57 subjects across all 45 word pairs (median = 37%, SD = 15%) and the proportion correct for each word pair across all subjects (median = 33%, SD = 26%). Because of the design of our study, these memorability measures correspond to the proportion correct for a word pair in a particular sequence in the list that every subject in our study observed. The distributions of word pair memorabilities and subject memory abilities are shown in Fig. 2. To gain a further sense of how much of the variance in memory performance was explained by the task itself, prior to examining neural activity, we fit a one-parameter item-response theory (IRT) model and examined its predictive performance. Because some subjects performed the final recall test online, we included a person-level parameter that allowed performance to vary between groups (79). Specifically, we fit the model $P(r_{ws} = 1|s, w) = \text{logit}^{-1}(\theta_s + \delta \mathbb{1}_{l_s=Online} + b_w)$ with $\theta_s \sim N(0, \sigma_\theta)$, where w indexes Lithuanian–English word pairs, s indexes subjects, r_{ws} is the response (correct or incorrect recall) of each subject to each word pair, l_s is a variable that indicates the location in which a subject took the recall test (Online or in Lab), δ represents the difference in performance between the groups, and b_w and θ_s represent the latent word memorability and subject ability parameters, respectively. This model estimates one parameter for each subject which captures their overall ability and one parameter per word-pair which captures its overall difficulty

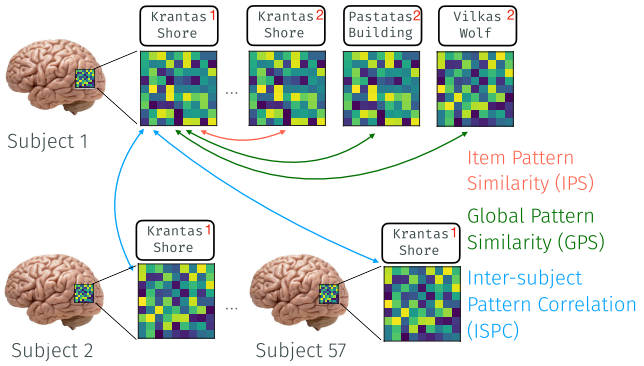

Fig. 3. Methods for constructing the three types of multivariate neural features from fMRI signals recorded during encoding. Each word pair in the experiment was presented five times, and red numbers in the *Upper Right* corner indicate the repetition number (but were not present during the actual experiment). Item Pattern Similarity (**IPS**) is computed by comparing two presentations of the same word pair. Global Pattern Similarity (**GPS**) is computed as the mean similarity between a given word pair presentation and presentations of all 44 other word pairs from a different repetition (shown here: repetitions 1 and 2). All repetitions of an item took place on separate scanner runs, so only including presentations from different runs ensured that similarity was purely driven by the response to the item and not, e.g., shared scanner noise. Inter-Subject Pattern Correlation (**ISPC**) is computed by comparing a single presentation of a word pair for one subject to the same presentation of that word pair for all other subjects and averaging the similarities.

in the population our subjects were drawn from.[‡] We fit this model repeatedly, leaving one subject's data out and evaluating their predictive performance based on the area under the receiver operating characteristic curve (AUC 82–84), computed for the held-out subject. This IRT model achieves an average held-out AUC of .72.

Neuroimaging Analyses. As validation of our neural recordings, we first created standard subsequent memory maps, across all repetitions individually and averaged (*SI Appendix*, Fig. S5–S10). We next attempted to predict memory performance on each individual word pair for each subject from the four types of neuroimaging signals described above. Fig. 3 shows how we computed each of the multivariate measures from the trial-evoked neural signals recorded during the task. We use a standard ridge regularized logistic regression[§] approach (86), which has commonly been used for studying the neural predictors of memory, e.g., refs. 42, 87–89, and fit separate models for each of the four features, allowing us to interpret which of the features were associated with successful memory prediction, both without and with adjustments for the memorability of individual word pairs within the context of the list. We focus on linear models here due to their widespread use in the literature and leave evaluating possible nonlinear relationships, such as interactions across time and space, to future work. Our approach of combining regularized machine learning models with statistical adjustment for confounders is related to the value-added prediction method of Reiss et al. (90), leveraging the Potter (91) approach to adjustment. Conceptually, however, we believe that the approach we advocate here is much more straightforward.

To compute the features, we first estimated individual voxel BOLD activation associated with the onset of each study trial with a general linear model (GLM) using the least squares-single approach described in ref. 92. Given these maps of activation, we could then compute the four features associated with each trial presentation. Ideally, because regularization downweights features that are less useful for prediction, we would include every feature in a single model and infer the optimal predictive model. However, in practice, it is well known that including more irrelevant features requires more regularization, limiting predictive performance (93). In addition, it has long been acknowledged that smoothing can improve the performance of MVPA analyses (94). This is especially true in the across-subject prediction setting where individual voxels may not be perfectly aligned (95, 96). Finally, taking averages (or weighting a number of features equally) can have good statistical properties over learning weights for individual features, especially in small datasets, e.g., robustness to noise (97, 98). We therefore describe several approaches below for models that vary in their number of features as well as amount of aggregation prior to computing features, which allows us to test the sensitivity of our approach to these two concerns and strike a balance between model flexibility and sensitivity to small effects.

Because each word pair was presented five times, we considered several approaches to combining the five measurements to predict a single recall test. One approach is to simply include all features, one for each study trial, and allow the statistical model to infer their relative importance from the training data, as it may be that measurements taken early on or closer to the end of the study session might be more relevant than others. Another approach is to take the mean of each feature across the five presentations. This approach has been used in previous studies with multiple presentations of the same item, e.g., ref. 15.

Multivariate neuroimaging signals necessarily involve computing the relationship between several measurements, typically neighboring ones, e.g., grouped via an anatomical map into predefined brain regions of interest or using a searchlight analysis (94). In order to make our models using univariate activity more comparable in terms of number of features, as well as make them more robust to across-subject deviations, we aggregate voxel activation at the region of interest (ROI) level as well. In our most flexible version of the model, we computed each multivariate feature in each of 100 cortical ROIs from a parcellation of the brain defined by **Schaefer** et al. (99) (Fig. 4*A*). In order to restrict the number of irrelevant features, we test models using only a subset of **targeted** ROIs, selected to reflect the domain knowledge about regions likely to be involved in memory encoding in a verbal learning task (*Methods* and Fig. 4*B* for the exact definitions). Due to the central role the **hippocampus** is thought to play in memory formation (100–102) and because of the difficulty in recording precise fMRI BOLD signal from the hippocampus when using standard imaging sequences for targeting the whole brain (103), we test several models including only features computed in the hippocampus. The separate set means that the smaller signal-to-noise ratio in hippocampus BOLD will not prevent its inclusion in the penalized regression models we use for prediction. We test models that include the average activity, pattern similarity, and global pattern similarity in each individual's **whole** hippocampus and also hippocampal **subparts** (left and right, posterior, medial, and anterior). In addition, we also select the voxels that are included in all participants' anatomically defined hippocampus which allows us to test a classifier based on individual hippocampal **voxel** activity. Finally, we can define a **whole** hippocampus intersubject pattern correlation feature based on these overlapping voxels (Fig. 4*C*).

[‡]We initially tested a model that included random effects allowing for differential item functioning between the group that took the test online and the group that took the test in the lab (80, 81). However, we found no evidence for differential item functioning (*SI Appendix*) and therefore did not explore this model further.

[§]Also known as L2-regularized logistic regression (85).

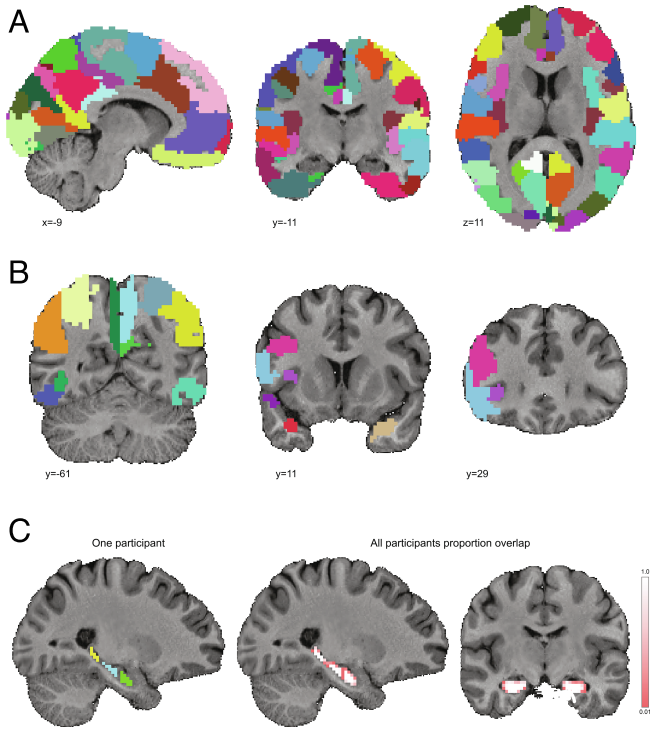

Fig. 4. Example location (MNI) and extent of ROIs used for predictions. (*A*) Ref. 99 100 ROI parcellation; (*B*) Example of the targeted ROIs; (*C*) Anatomically defined anterior, mid, and posterior hippocampal ROIs for one participant (*Left*) as well as the proportion of participants having a particular voxel assigned to one of the three hippocampal ROIs (*Middle* and *Right*).

As a final strategy for decreasing sensitivity to errors in voxel alignment, we follow a **whole-brain** strategy, inspired by ref. 88, where we treat the average activity in an ROI (using the same set of ref. 99 cortical ROIs as above) as a "voxel" for the purpose of computing similarity across item presentations in multivariate features.

Subsequent memory models. We first test the ability of these features to predict in a standard subsequent memory setting, that is, predicting memory performance only from features of the neuroimaging signals. These models have the form $P(r_{ws} = 1|s, w, \hat{\theta}, \mathbf{X}) = \text{logit}^{-1}(\hat{\alpha} + \hat{\theta}_s + \hat{\delta}\mathbb{1}_{l_s=Online})$, where w indexes a word pair, s indexes a subject, and $\mathbf{X}$ is one of the sets of neuroimaging signals defined above, normalized within subject. Because this model was fit across subjects, we also included subject-level intercepts to account for overall subject differences in recall performance. To accomplish this, we used the estimates of subject ability $\hat{\alpha} + \hat{\theta}_s + \hat{\delta}\mathbb{1}_{l_s=Online}$ from a random-effects model, $P(r_{ws} = 1|s, w) = \text{logit}^{-1}(\alpha + \theta_s + \delta\mathbb{1}_{l_s=Online})$ with $\theta_s \sim N(0, \sigma_\theta)$. We estimate the predictive power of these models with leave-one-subject-out cross-validation (105), using the area under the ROC curve as a performance metric. Within each training set, we fit these models using a ridge penalty on the β parameters and choose the amount of regularization using ten-fold cross-validation. To evaluate the models' generalization performance, we conduct statistical tests on the held-out AUCs. For the standard models, we compute a one-sample t-test, comparing the model's performance to the AUC of a random guessing model (.5). In leave-one(-subject)-out cross-validation, the distribution of AUC scores (or any other metric) across holdout sets will in general be correlated because the training set for the classifiers will be largely the same. The independence assumptions of the t-test are therefore violated (106). To remedy this, we use a permutation test to estimate the empirical null distribution of paired t-statistics when there is no relationship between the fMRI features and recall (107–110).

In Fig. 5, we plot the mean and SE of the AUC estimates in held-out subjects. We show results using both the mean feature across all repetitions as well as using all study block and study-block pairs. Several of our feature/ROI combinations, when aggregated at the mean level, predict memory significantly above chance (.5) based on a permutation test. These include ISPC, GPS, and activity models using all Schaefer ROIs as well as when using only the set of targeted ROIs (M = .56 to .63, permuted Ps < .05). Results are overall similar when including each presentation or pair of presentations separately except that the ISPC models and the Targeted GPS model are no longer significant. Finally, a model that uses all available features from all models tested predicts significantly above chance. Because we are testing so many models, it may make sense to control the overall false discovery rate across all models tested (104). When we calculate the adjusted q values, the Schaefer GPS model including all pairs of repetitions and the Targeted mean GPS model are no longer significant.

Overall, this suggests that several features we tested are able to successfully predict memory across subjects and would normally be classified as subsequent memory effects. In particular, when averaging over all study blocks and including all cortical ROIs, ISPC, GPS, and univariate activity could all successfully predict memory in a held-out subject. While most MVPA analyses of memory perform within-subject prediction on new sessions, this is a demonstration of across-subject prediction of recall from neuroimaging measures of encoding. This demonstrates that these features of the neuroimaging signal contain a significant amount of information about memory performance. Next, we

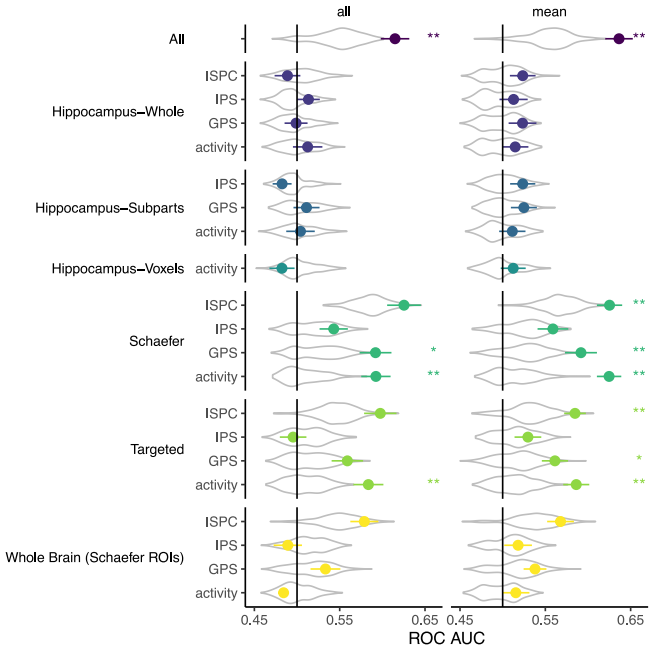

Fig. 5. Classifier performance based on the Standard Subsequent Memory Model. Each combination of ROIs, features, and time-point treatment is plotted separately with definitions of terms found in *Methods*. IPS = Item Pattern Similarity, GPS = Global Pattern Similarity, ISPC = Intersubject Pattern Correlation. The black line indicates chance AUC (.5), and statistical tests are compared with this baseline. * = $P < .05$ based on a permutation t-test, ** = $q < .05$ after false discovery rate (FDR, 104). Error bars reflect the unadjusted SEM. Violin plots reflect permutation null distributions.

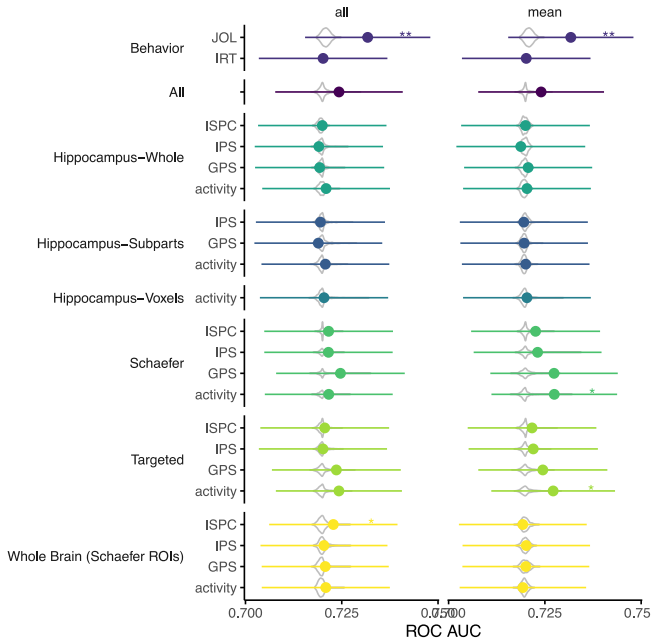

Fig. 6. Classifier performance based on the ICEA Subsequent Memory Model. Each combination of ROIs, features, and time-point treatment is plotted separately with definitions of terms found in *Methods*. IPS = Item Pattern Similarity, GPS = Global Pattern Similarity, ISPC = Intersubject Pattern Correlation, JOL = Judgment of Learning, IRT = Item Response Theory model. JOL and IRT are plotted in each column for comparison although they do not differ across columns. * = $P < .05$ based on a permutation t-test, ** = $q < .05$ after an FDR (104) adjustment. Error bars reflect the unadjusted SE of the mean. Violin plots reflect permutation null distributions.

will test whether there is evidence that the signals reflect activity that is causally involved in encoding or downstream from such activity.

ICEA subsequent memory models. The IRT model described above estimates the expected performance for each item in the list without considering a subject's neural signals, essentially estimating the total effect of potential confounding variables (mean AUC = .72). Given the performance of that model, we can see that the potential confounding factors in a standard memory task already account for a significant amount of the variance in memory behavior.

To estimate ICEA-related subsequent memory effects, we then combine this IRT model with a model for predicting memory from the neuroimaging features. If this model has a stronger relationship with memory performance than the IRT model alone, we argue that this provides evidence that the neuroimaging features reflect ICEA. To do so, we fit a ridge regularized logistic regression that includes a fixed intercept (or offset), which is the linear predictor for each subject and word pair that was estimated previously in the IRT model, i.e., $\hat{\theta}_s + \hat{\delta}\mathbb{1}_{l_s=Online} + \hat{b}_w$. The full model we estimate is therefore: $P(r_{ws} = 1|s, w, \hat{\theta}, \hat{\mathbf{b}}, \mathbf{X}) = \text{logit}^{-1}((\hat{\theta}_s + \hat{\delta}\mathbb{1}_{l_s=Online} + \hat{\mathbf{b}}_w) + \beta\mathbf{x}_{ws})$, where $\mathbf{x}_{ws}$ is a vector of neural features (e.g., average activity in several ROIs) for subject s on word pair w. The distribution of held-out AUCs for each model is plotted in Fig. 6. We compare each model's held-out AUC performance to the IRT model using a paired t-test and using permutations to construct the null distribution. Among these models, only the Targeted (M = .727) and Schaefer (M = .727) Activity models using the mean of the study blocks and the Whole-Brain ISPC model (M = .723) using all study blocks perform significantly better than the IRT model (permuted Ps $< .05$). However, after false discovery rate adjustment, none of the q values are less than .05. Thus, there is no reliable statistical evidence that any of the fMRI features we included in our models are reliably measuring ICEA. In *SI Appendix*, Figs. S14 and S15, we examine the same set of models using only features from a single study trial or study trial pair, finding qualitatively similar results there as well.

Judgments of learning. We now test whether the behavioral judgments of learning (**JOL**) ratings appear to reflect ICEA. We can use the same modeling approach we employed above, predicting memory from the judgments of learning while adjusting for confounding variables using the parameters from the IRT model. The model including judgments of learning did predict significantly better than the IRT model, improving the AUC on average by .0116 ($t_{(56)} = 6.84$, $P < .002$).[¶] Assuming that the JOLs are a behavioral readout of neural activity, this implies that the underlying activity is ICEA by the definition adopted earlier. However, given our permutation approach to hypothesis testing, we cannot compare the magnitude of the JOL predictive power to that of the fMRI signals. Therefore, we do not claim that the underlying activity that produces a JOL ICEA is not causally stronger than that underlying the fMRI signals, but merely that the JOL ICEA signal is reliable enough to be detected.

Discussion

We found that several univariate and multivariate features of fMRI data proposed in the subsequent memory literature (2, 3) could predict cued recall over a long delay in a verbal paired-associates task. We also demonstrate that Intersubject Pattern Correlation (ISPC) is a predictor of memory in a task not using video stimuli and that Global Pattern Similarity (GPS) outside of the hippocampus predicts memory in a recall task.

However, like in many memory experiments, a significant amount of the variation in the probability of a subject remembering a particular item was explained by the average performance in the population for the specific item in its particular experimental context. A common perspective on subsequent memory effects is that they index "the depth of encoding of the to-be-remembered stimuli" and "determine the efficacy of memory encoding" (18), suggesting that the underlying activity is causally involved in memory encoding. But because the effects that lead an item to be more likely to be remembered on average might also drive neural activity, the existence of variation in memorability across items confounds a causal interpretation of subsequent memory effects based simply on raw associations. We therefore tested to what extent memory-related fMRI signals were related to variation in recall after adjusting for possible confounding variables such as item memorability (29), serial position (43), and list composition (45). Our analyses, based on comparing predictive models using only behavior with models using behavior and fMRI data combined, showed that the subsequent memory effects we observed were highly correlated with population memorability and did not consistently improve predictions across subjects once the predictive model included the average memorability of an item in a particular context in the population, estimated via a psychometric model. This suggests that the causal interpretation of the activity measured by subsequent memory effects may be less warranted than typically assumed. At least in our data, these analyses suggest that objective

[¶] This was the smallest possible p-value in our permutation test given that we ran 500 samples. The true P-value may be much smaller, but we do not have the precision to report it.

characteristics of target items may drive a significant portion of the subsequent memory effect. In addition, to the extent that we can rely on ICEA results that were nominally significant, although not at standard levels after false discovery rate adjustment, to indicate possible interesting signals to investigate in future work, the standard subsequent memory effect was not particularly correlated with ICEA results. Indeed one of the significant ICEA signals was the whole-brain intersubject pattern correlation, a signal that was not a significant subsequent memory effect and one that we do not know of any prior work on in the existing literature on cognitive neuroscience of memory.

Of course, it is entirely possible that our inability to derive a statistically reliable signal of plausible causal encoding activity is simply due to higher variability in our behavioral and neuroimaging data than the typical subsequent memory study. In our experimental design, we did not explicitly manipulate task demands, nor impose a strategy on subjects. In addition, the nature of our task (long delays, associative memory, and novel stimuli) may have prevented some processes, perhaps those causally involved in memory formation, from being involved that may typically be measured by subsequent memory effects. However, we argue that the strength of the results using the standard subsequent memory approach, without accounting for confounders, and the results of the ICEA subsequent memory approach when including the judgments of learning as a signal of mnemonic status during a temporally distal period place limits on these alternative explanations. In the following, we discuss several interpretations of these analyses and their implications for future studies of subsequent memory. We additionally provide further detail on the relationship of the present study to past work in *SI Appendix*.

One explanation is that much of the variability in memory that could be predicted at encoding is explained by the psychometric model already. For the fMRI features to predict when controlling for the confounding variables, they would have to be reliably indexing causal encoding processes that contradict the prediction from a model based on population averages (e.g., a failure to pay attention to a particular item). Some might argue that these occurrences are relatively rare, especially if participants are focused on the task, therefore reducing our ability to observe ICEA. In addition, due to the long delay, there is the potential for a lot of variability in memory accessibility between encoding and retrieval due to events that happen in between, limiting the predictability of memory in our task. However, the performance of the model including Judgments of Learning shows that there is a signal (that is in fact consciously available to subjects) that consistently predicts subsequent memory even across items with the same population memorability. This model gives a lower bound on what is explainable near the time of encoding and very distant from the time of the retrieval test.

A second possible interpretation is that there is too much noise in the fMRI data to hope to find signals comparable to a judgment of learning in a data set of our size. However, at 57 subjects, our data set is approximately 2 to 3 times the size of most data sets in the subsequent memory literature, although we use fewer items than typical studies using recognition memory. In *SI Appendix*, we conduct a post hoc power analysis and compare to other study designs in the literature, showing that our experiment likely had greater power than most other subsequent memory studies. This indicates that future work trying to isolate ICEA may require large data sets and tightly controlled designs.

A third possible interpretation is that there are indeed signals of ICEA that exist in fMRI data, but they are highly variable across people, perhaps due to a lack of precise voxel alignment or differently shaped brains causing variation in the signal-to-noise characteristics. If this were the case, it may be that our attempt to do across-subject prediction was doomed from the start. However, we point to the success of our associative subsequent memory models (without adjusting for confounding variables) as evidence that these issues do not prevent any across-subject prediction in our data. In addition, several of our features made efforts to circumvent this difficulty by a) aggregating data at higher levels (e.g., the whole-brain analyses treating each ROI as a "voxel"), b) using hippocampal ROI definitions based on individual subjects' anatomy, and c) using multivariate measures like pattern similarity that are less sensitive to perfect alignment across subjects. However, none of these features seem to have succeeded in improving classifier accuracy significantly beyond the IRT model.

A fourth possible interpretation is that the actual cognitive strategies used for learning (especially in an undirected, intentional encoding task like ours) are highly variable across people. For instance, it may be that high-memorability word pairs are easier to sound out, resulting in greater activity in phonological regions on average. But if sounding out is an effective strategy for only some people, activity in a particular phonological region may not consistently predict memory. Across-subject prediction differs from the usual two-level within-subject paradigm that is typical of MVPA studies in that it is more sensitive to small-magnitude, low-variability effects but less sensitive to high-magnitude, high-variability effects (95). If the measurement of encoding processes is more like the latter scenario, this suggests that future studies of subsequent memory will require larger data sets that allow for understanding groups of subjects who use similar encoding strategies.

A fifth possible interpretation is that certain aspects of our design, such as the inclusion of multiple study opportunities and a judgment of learning task, diminished the relationship between ICEA during any one word pair presentation and future memory, relative to other more typical SME designs. As noted above, we used multiple repetitions to allow for testing of IPS, an important and theoretically motivated predictor of subsequent memory (15), as well as increasing the educational relevance of our study. However, it is possible that repetitions decreased the effect of individual moment-to-moment fluctuations in encoding quality on subsequent memory and increased the effect of item and task effects. Therefore, future studies seeking to isolate ICEA may prefer to not investigate IPS and use single-item presentations, which may increase the magnitude of the ICEA signal in fMRI. Indeed, there may be reasons to think that stability of item representations across presentations is related to properties of the items (70) and is therefore unlikely to reflect ICEA.

It is similarly possible that judgments of learning provided an additional study or retrieval practice (111, 112) opportunity although effects may be quite small (113, 114), thus decreasing the relevance of ICEA during the previous study trials on future memory performance. However, we do not think that the inclusion of these JOLs was a major concern for two reasons. First, we do still observe the SME using the *Standard Subsequent Memory Model*, suggesting that the judgment of learning trial did not completely swamp all of the variance in memory and prevent the prior fMRI signals from predicting memory. Second, we use a long (by subsequent memory effect standards) study-test delay of 72 h. In a more typical study where the test came shortly after the study period, a judgment of learning trial in the intervening period could have a very large effect on the mnemonic status at

test time. However, in our task, subjects have many uncontrolled experiences, in addition to the judgment of learning, in which they may reencounter words they saw in the task, similarly providing a retrieval practice opportunity. Therefore, we believe that the effect of the additional judgment of learning is likely to be relatively small. In fact, it is fairly amazing that the judgment of learning reads out a causal signal that contributes to memory 3 d later, despite the potential for intervening reminders. However, future studies investigating ICEA may choose not to include JOLs as a way to increase similarity to previous studies.

Overall, it is certainly possible that our study was limited by a low neuroimaging signal-to-noise ratio and high across-subject variability in both the fMRI BOLD measurement and the cognitive processes involved in encoding themselves. However, we argue that the success of the JOLs as a predictor of memory after adjusting for confounding and the existence of successful memory prediction from fMRI in the absence of item and task variables suggest that this cannot be the complete story.

Conclusion

In the 40 y since the first subsequent memory effect was reported (4), much has been learned about the neural signals associated with memory. A vast literature maps the various univariate and multivariate signals across the brain, along with proposed cognitive computations being implemented by the underlying neural activity. In addition, many signals have supporting evidence from animal and lesion studies. This paper contributes to this literature by documenting the existence of several subsequent memory effects in a novel language-learning task. We show that many features proposed to be relevant for other types of memory tasks such as Intersubject Pattern Correlation and Global Pattern Similarity also apply here. However, the exact interpretation of these subsequent memory effects remains ambiguous, and the original hope of identifying the causal mechanisms of memory encoding has not been fully realized. Here, we defined a framework for making progress toward identifying (Indicators of) Causal Encoding Activity from neuroimaging signals. Based on this formulation, we attempted to precisely control for many exogenous effects including stimulus, presentation order, and overall list context effects that have been documented to influence the probability of recall in order to see whether we can extract signals from fMRI that are still associated with subsequent memory performance, reflecting activity that is either causally involved in memory encoding or downstream from such activity. While we were not able to identify such signals, we suggest that future work leverage the techniques of this paper, i.e., classifier analyses and tight control of memorability, with larger datasets, different tasks, and new features of the fMRI signal. By controlling even more aspects of the task, future work can remove additional potential confounds of ICEA. In several ways, our own design could be improved, for instance, by using items with more similar memorability scores (so that more variance can be explained by the neural signals), by including a movie-watching portion of the scanner task that would allow for the use of better alignment tools (96, 115) or by using a task such as recognition memory that would allow for more trials to be collected per participant. Despite these qualifications, we think that these results suggest that interpretations of classic subsequent memory results may not be as straightforward as commonly assumed and lay out key questions for cognitive neuroscientists of memory to address in the future. Similar to recent work in other fields of neuroscience, explicit causal inference can overturn common interpretations of correlational results (116). We do not conclude that there are no memory-related causal signals available to noninvasive observation. Instead, we want to highlight the message that appropriate accounting for various stimulus and contextual factors can facilitate the identification of these signals. Indeed, we fully expect that people will identify such markers of memory formation (or have already done so) and will be able to causally interact with memory formation in the near future.

Materials and Methods

Participants. Sixty-nine participants were recruited using electronic advertisements hosted by New York University and accessible to the broader community. We obtained written informed consent prior to conducting the study. This study was approved by the New York University Institutional Review Board. All participants self-reported to be between 18 and 35 y old, had normal or corrected-to-normal vision, spoke English fluently, and did not speak Lithuanian or a related language.

Eleven participants' data were excluded from analysis based on MRI data issues: two participants' MRI sessions were conducted with incorrect scanning parameters; six participants were excluded based on excessive motion and other image quality issues; three participants requested early termination of the experiment due to discomfort in the scanner (and did not complete the recall test). We also excluded one of the participants who recorded 100% correct in the recall test since their held-out AUC would be undefined, resulting in a final data set of 57 participants

Behavioral Task. Based on a normed set of Lithuanian–English words, we selected 45 translation pairs with a range of difficulties (75) for use in our cued recall task, also described in Fig. 1. All participants first completed a study phase inside the fMRI scanner where they saw the translation pairs presented one at a time for 4 s each with a variable duration intertrial interval (randomly chosen between 4 and 16 s). During this intertrial interval, in order to prevent rehearsal, participants made judgments about whether each of a series of numbers was odd or even.

Words were presented on a computer screen with the Lithuanian word at the top of the screen and the English translation underneath. Each word pair was presented five times, and no pair was presented for the nth repetition until all words had $n-1$ presentations. Importantly, and in contrast to many psychology studies on the subsequent memory effect, all participants see the same sequence of study items. At the expense of generalizability, this allowed us to precisely quantify the average recall performance for a word pair in a specific context and account for any interactions between items. The order of the word pairs was selected as follows: The 45 words were grouped into groups of five words each. On each repetition, the order of the five words was shuffled, but all five words appeared before the next group of five were presented. Using this approach, we sampled a single order of item presentations that was then used for all 69 subjects in our study. Immediately following the study session, participants completed a second section outside the scanner that involved making judgments of learning JOLs (78): For each pair, participants were presented with the Lithuanian word and English word and used the computer mouse to indicate on a scale of 0 to 100 how likely they were to remember the association in 3 d. Participants had up to 12 s to respond, and the response was coded as missing if this deadline was not met. (In practice, all participants gave all JOL ratings within the allotted time.)

Participants were then asked to complete a recall test approximately 72 after the initial scan session. The first 44 of the subjects completed this in the lab, and the final 13 subjects completed the second session at home, via an online version of the task to ease the task of scheduling sessions with a 72-h delay between. Overall task performance varied slightly but statistically significantly between the sample that completed the recall task in the lab compared to those who completed it online (*SI Appendix*). This difference may have been due to a number of factors including that the data were collected at a different time and that some Internet capability was required to complete the task. However, word memorability did not differ across the two test modalities, so we proceed using similar models for both sets of subjects. Participants saw a Lithuanian word presented on the screen and had to type the associated English word. A trial was coded as correct if participants typed the correct English word (allowing for

typographic errors) and all other responses were incorrect. Like the judgments of learning, this task was completed outside the scanner.

fMRI Data Acquisition. MRI data were acquired on a 3 Tesla Siemens Prisma scanner. Anatomical images were collected using a T1-weighted MPRAGE high-resolution sequence (0.8-mm isotropic voxels). Five runs of whole-brain BOLD functional data were acquired using an EPI sequence (2.5-mm isotropic voxels; TR = 1 s; TE = 35 ms; multiband factor = 4; phase encoding: anterior–posterior). An additional pair of nonaccelerated EPI scans with opposing forward (anterior–posterior) and reverse (posterior–anterior) phase encoding relative to the primary functional data was collected for distortion correction during preprocessing (2.5-mm isotropic voxels; TR = 4.496; TE = 45.6). The slice position and orientation for these scans matched those of the BOLD data collected during the behavioral task.

fMRI Preprocessing. Results included in this manuscript come from preprocessing performed using fMRIPrep 1.2.6-1 (117), which is based on Nipype 1.1.7 (118).

Anatomical data preprocessing. The T1w anatomical scans were corrected for intensity nonuniformity (INU) using N4BiasFieldCorrection (ANTs 2.2.0, 119). A T1w-reference map was computed (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1, 120). The T1w-reference was then skull-stripped using antsBrainExtraction.sh (ANTs 2.2.0), using OASIS as the target template. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, 121), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray matter of Mindboggle (122). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (123) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0, 124), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9 125).

Functional data preprocessing. For each of the five BOLD runs collected per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. A deformation field to correct for susceptibility distortions was estimated based on two echo-planar imaging (EPI) references with opposing phase-encoding directions, using 3dQwarp (AFNI 20160207, 126). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate coregistration with the anatomical reference.

The BOLD reference was then coregistered to the T1w reference using bbregister (FreeSurfer), which implements boundary-based registration (127). Coregistration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, 128). The BOLD time series were resampled onto their original, native space by applying a single, composite transform to correct for head motion and susceptibility distortions. These resampled BOLD time series will be referred to as preprocessed BOLD in original space or just preprocessed BOLD. The BOLD time series were resampled to MNI152NLin2009cAsym standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space.

Several confounding time series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS, and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by 129). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, 130). Principal components are estimated after high-pass filtering the preprocessed BOLD time series (using a discrete cosine filter with 128-s cutoff) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures that it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e., head-motion transform matrices, susceptibility distortion correction, and coregistrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (131). Nongridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer). Preprocessed BOLD data were smoothed to 6-mm FWHM using AFNI 3dBlurToFWHM.[#]

Regions of interest (ROIs).

Gray matter and hippocampus anatomical masks Analysis of the functional data was restricted to voxels encompassed by a binary mask computed from the probabilistic gray matter estimates generated during the FreeSurfer anatomical preprocessing combined with anatomically defined voxels in individual hippocampi. For each participant, the probabilistic gray matter mask was binarized with a threshold of 0.2 and intersected with bilateral hippocampal voxels as estimated using FSL FLIRT implemented in Nipype (128, 132). The combined gray matter/hippocampus mask was resliced to functional resolution, smoothed with an 8-mm FWHM kernel to be liberal in voxel inclusion (we opted to include some potentially nongray matter voxels that could be ignored in downstream analyses rather than being overly strict and excluding relevant signal) and then rebinarized with a threshold of 0.2.

Whole-brain cortical ROIs For whole-brain cortical analysis, voxels were aggregated into ROIs provided by ref. 99. To strike a balance between regional specificity and total number of cortical ROIs, we used the intersection of the gray matter mask defined during preprocessing and the 100-parcel, 7-network atlas from ref. 99[‖] after reslicing to 2.5 mm^3.

Targeted ROIs We defined a targeted set of nineteen ROIs on the basis of prior expectations of these regions' potential engagement in a learning task involving verbal materials. We selected left and right perirhinal cortex (PRC) and anterior temporal lobe (ATL) ROIs due to putative roles for these regions in processing conceptual information (133). In addition, we selected regions highly likely to be involved in memory formation (2) and in processing concrete nouns (a category which includes all of the word pairs in our stimulus set) (134). These ROIs included two separate areas in the left anterior insula, left and right lateral occipital cortex, left and right dorsal parietal, left and right lateral parietal, left and right medial parietal (precuneus), left and right retrosplenial cortex, an upper and a lower portion of the left ventrolateral prefrontal cortex (VLPFC), and visual word form area (VWFA; left hemisphere only). This set of targeted regions listed above were defined from a variety of sources. The PRC regions were defined using the anterior portion (MNI $y > -21$) of the PRC ROIs provided by ref. 135 as downloaded from Neurovault[**] which overlapped with the PRC center coordinates reported by ref. 133. The ATL ROIs were defined as the ref. 99 ROIs encompassing the peak voxels reported by ref. 133 (99, 600 parcellation ROIs 191 and 492).

The remaining targeted ROIs were defined by identifying voxel groupings from ref. 99 which encompassed other cortical regions expected to be involved in memory formation and/or processing of verbal materials.

Single trial activation estimates. Individual voxel BOLD activation on each study trial was estimated with a general linear model (GLM) as implemented

[#] https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dBlurToFWHM.html.

[‖] https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal.

[**] https://neurovault.org/collections/3731/; https://identifiers.org/neurovault.collection:3731.

in SPM12 (136)[††] using the least squares-single approach described in ref. 92. For each study trial, a voxel-wise regression was carried out with one predictor containing a 4-s boxcar covering the time points corresponding to the to-be-estimated trial, another predictor containing four-second boxcars covering the time points for all of the remaining trials, and fourteen additional predictors containing a variety of other potential confounds and nuisance signals as estimated by fMRIPrep during preprocessing (129). These confounds and nuisance signals are related to head motion estimates and BOLD fluctuations of no interest estimated from voxels outside of the gray matter. The full set of nuisance regressors included the following: six parameters representing estimated translational and rotational head motion in each of three dimensions; two regressors with BOLD time courses estimated from cerebrospinal fluid (CSF) and white matter (WM) masks; and the top six components estimated from the aCompCor procedure which estimates temporal principal components from a combined nuisance mask as described in ref. 130 and implemented in fMRIPrep. To clarify, these confounds are confounds of the BOLD response on a particular trial, and the purpose of regressing them out is to obtain closer estimates of the true signal. In contrast, confounds we mention above are confounds of the causal relationship between a neural signal and memory performance. Prior to estimation, the vectors of single-trial and all-other trials were convolved with the canonical SPM double gamma hemodynamic response function (HRF) along with its temporal and dispersion derivatives (137). We then estimated regression coefficients for each voxel using ordinary least squares, and the procedure was repeated for each study trial for each participant. The resulting coefficients for the single-trial predictors were taken as summaries for the activation for each voxel on each trial.

fMRI feature estimates. After initial fMRI preprocessing, additional steps to create the fMRI features for predictive models and reshape data were conducted using custom code in python 3.7.7 using the packages pandas (138) and numpy (139) and R 4.0.2 using the package collections tidyverse (140) and tidymodels (141).

Multivariate patterns were derived from the single trial estimates for each voxel in a particular ROI. While several distance metrics have been proposed (142, 143), in order to remain faithful to the proposed features, we used Pearson's *r*, which is the metric used by nearly every paper in the subsequent memory literature, e.g., refs. 15, 68 and 66. In addition, for inclusion in the model, all of the features are Fisher-*z* transformed (144), which is often used when aggregating correlations since it makes the sampling distribution approximately normal, which is slightly better for use in the regularized regression models we describe in the following section.

Predictive Modeling.

IRT model. We estimate this model using maximum likelihood, as implemented in the R package lme4 (145). Like the models involving fMRI data, these models were trained in a leave-one-subject-out manner. When predictions are made using these models for a held-out subject, θ_S was set to 0, where *s* indexed the held-out subject.

Regularized regression models. To estimate the models, we use cyclical coordinate descent as implemented in the glmnet package (146) in R. We estimate the predictive power of our models with leave-one-subject-out cross-validation (105), using the area under the ROC curve as a performance metric. We choose the size of the penalty on the β parameters (λ) using nested 10-fold cross-validation. We choose a set of λ hyperparameters to try using the default algorithm in glmnet: first λ_{max} is chosen such that a model fit with that λ on all of the data will have all parameter estimates near 0, using a formula from (146). Then, 100 parameters are chosen on a log scale from λ_{max} to $\frac{\lambda_{max}}{10000}$. The model is estimated for each λ, and the λ that maximizes the log likelihood of the held-out fold is chosen for evaluation. All of our fMRI-derived features are standardized (by their mean and variance in the training fold) before being included in the model.

Evaluation. To evaluate the models' generalization performance, we conduct statistical tests on the held-out AUCs. For the standard models, we compute a one-sample *t*-test, comparing the model's performance to the AUC of a random guessing model (.5). AUC is a useful metric here because it is not affected by the base rates of remembering and forgetting. For the main test, we compare the ICEA models to the behavioral model using a paired *t*-test. To give the strongest possible chance to the MRI features, for all models, we use a one-sided test of whether the AUC in the ICEA model is greater. In leave-one(-subject)-out cross-validation, the distribution of AUC scores (or any other metric) across holdout sets will in general be correlated because the training set for the classifiers will be largely the same. The independence assumptions of the *t*-test are therefore violated (106). To remedy this, we use a permutation test to estimate the empirical null distribution of paired *t*-statistics when there is no relationship between the fMRI features and recall (107–110). To do so, we shuffle the pairing between the fMRI features and the Lithuanian–English word pair within subject. This allows us to maintain the correlation between fMRI features as well as the distribution of word-pair population memorabilities and subject recall abilities. Crucially, this also means that IRT does not change its predictions, so the residual variance to be explained remains the same across permuted datasets. When shuffling the data, we also maintain the structure of the sequence of items such that the same five items appear in each group of five serial positions of each study block. We created 500 permuted datasets, applied to the same classification models and computed a paired *t*-statistic for the permutation test.

Data, Materials, and Software Availability. Anonymized Behavioral and preprocessed fMRI data. Data have been deposited in OSF (osf.io/hrac5/).

ACKNOWLEDGMENTS. We thank Hong Yu Wang, Camille Gasser, and Steven Mikal for helpful assistance with stimulus development, scanning, and data processing. We additionally thank Daniel Schonhaut, Noa Herz, John Sakon, Michael Kahana, Joseph Halpern, Cate Hartley, and Eero Simoncelli for helpful comments on previous drafts of this manuscript and Rich Shiffrin, Anthony Wagner, and two anonymous reviewers for their insightful critiques and suggestions during the review process. This work was supported by NSF grant DRL-1631436 and seed funds from the NYU Dean for Science.

[††] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

1. K. A. Paller, A. D. Wagner, Observing the transformation of experience into memory. *Trends Cognit. Sci.* **6**, 93–102 (2002).
2. H. Kim, Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage* **54**, 2446–2461 (2011).
3. G. Xue, The neural representations underlying human episodic memory. *Trends Cognit. Sci.* **22**, 544–561 (2018).
4. T. F. Sanquist, J. W. Rohrbaugh, K. Syndulko, D. B. Lindsley, Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology* **17**, 568–576 (1980).
5. A. D. Wagner, Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281**, 1188–1191 (1998).
6. L. Davachi, J. P. Mitchell, A. D. Wagner, Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2157–2162 (2003).
7. G. Xue *et al*., Greater neural pattern similarity across repetitions is associated with better memory. *Science* **330**, 97–101 (2010).
8. M. R. Uncapher, M. D. Rugg, Selecting for memory? The influence of selective attention on the mnemonic binding of contextual information*J. Neurosci.* **29**, 8270–8279 (2009).
9. M. Aly, N. B. Turk-Browne, Attention promotes episodic encoding by stabilizing hippocampal representations. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E420–E429 (2016).
10. I. Charest, R. A. Kievit, T. W. Schmitz, D. Deca, N. Kriegeskorte, Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14565–14570 (2014).
11. M. T. R. van Kesteren, G. Fernandez, D. G. Norris, E. J. Hermans, Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7550–7555 (2010).
12. L. J. Jenkins, C. Ranganath, Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. *J. Neurosci.* **30**, 15558–15565 (2010).
13. L. J. Lohnas, L. Davachi, M. J. Kahana, Neural fatigue influences memory encoding in the human hippocampus. *Neuropsychologia* **143**, 107471 (2020).
14. M. D. Rugg *et al*., Item memory, context memory and the hippocampus: fMRI evidence. *Neuropsychologia* **50**, 3070–3079 (2012).
15. G. Xue *et al*., Greater neural pattern similarity across repetitions is associated with better memory. *Science* **330**, 97–101 (2010).
16. S. C. Berens, J. S. Horst, C. M. Bird, Cross-situational learning is supported by propose-but-verify hypothesis testing. *Curr. Biol.* **28**, 1132–1136.e5 (2018).
17. Y. Ezzyat *et al*., Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nat. Commun.* **9**, 365 (2018).

18. K. Fukuda, G. F. Woodman, Predicting and improving recognition memory using multiple electrophysiological signals in real time. *Psychol. Sci.* **26**, 1026–1037 (2015).
19. J. M. Montgomery, B. Nyhan, M. Torres, How conditioning on posttreatment variables can ruin your experiment and what to do about it: STOP conditioning on posttreatment variables in experiments. *Am. J. Polit. Sci.* **62**, 760–775 (2018).
20. P. R. Rosenbaum, The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. Ser. A (General)* **147**, 656 (1984).
21. S. Weichwald, J. Peters, Causality in cognitive neuroscience: Concepts, challenges, and distributional robustness. *J. Cognit. Neurosci.* **33**, 226–247 (2021). http://arxiv.org/abs/2002.06060.
22. M. P. Grosz, J. M. Rohrer, F. Thoemmes, The taboo against explicit causal inference in nonexperimental psychology. *Perspect. Psychol. Sci.* **15**, 1243–1255 (2020).
23. M. A. Hernán, The C-word: scientific euphemisms do not improve causal inference from observational data. *Am. J. Public Health* **108**, 616–619 (2018).
24. S. M. Stoke, Memory for onomatopes. *Pedagogical Seminary J. Genet. Psychol.* **36**, 594–596 (1929).
25. A. M. Gorman, Recognition memory for nouns as a function of abstractness and frequency. *J. Exp. Psychol.* **61**, 23–29 (1961).
26. A. Paivio, Learning of adjective-noun paired associates as a function of adjective-noun word order and noun abstractness. *Can. J. Psychol./Revue Can. de Psychol.* **17**, 370–379 (1963).
27. J. R. Binder, R. H. Desai, W. W. Graves, L. L. Conant, Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
28. J. Wang, J. A. Conder, D. N. Blitzer, S. V. Shinkareva, Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* **31**, 1459–1468 (2010).
29. W. A. Bainbridge, "Memorability: How what we see influences what we remember" in *Psychology of Learning and Motivation* (Elsevier, 2019), vol. 70, pp. 1–27.
30. C. L. Baldwin, R. S. Runkle, Biohazards symbol: Development of a biological hazards warning signal. *Science* **158**, 264–265 (1967).
31. P. Isola, J. Xiao, A. Torralba, A. Oliva, "What makes an image memorable?" in *CVPR 2011* (IEEE, Colorado Springs, CO, USA, 2011), pp. 145–152.
32. J. Han *et al*., Learning computational models of video memorability from fMRI brain imaging. *IEEE Trans. Cybern.* **45**, 1692–1703 (2015).
33. D. C. Rubin, Memorability as a measure of processing: A unit analysis of prose and list learning. *J. Exp. Psychol.: Gen.* **114**, 213–238 (1985).
34. W. Xie, W.A Bainbridge, S. K. Inati, C. I. Baker, K. A. Zaghloul, Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nat. Hum. Behav.* **4**, 937–948 (2020), 10.1038/s41562-020-0901-2.
35. W. A. Bainbridge, D. D. Dilks, A. Oliva, Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage* **149**, 141–152 (2017).
36. W. A. Bainbridge, J. Rissman, Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Sci. Rep.* **8**, 8679 (2018).
37. J. D. Angrist, J. S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, 2009). OCLC: ocn231586808.
38. W. G. Cochran, D. B. Rubin, Controlling bias in observational studies: A review. *Sankhya: Indian J. Stat. Ser. A (1961–2002)* **35**, 417–446 (1973).
39. J. Hausman, W. Taylor, Panel data and unobservable individual effects. *J. Econ.* **16**, 155 (1981).
40. J. Pearl, Causal diagrams for empirical research. *Biometrika* **82**, 669–688 (1995).
41. C. Winship, S. L. Morgan, The estimation of causal effects from observational data. *Ann. Rev. Sociol.* **25**, 659–706 (1999).
42. C. T. Weidemann, M. J. Kahana, Neural measures of subsequent memory reflect endogenous variability in cognitive function. *J. Exp. Psychol.: Learn. Mem. Cognit.* **47**, 641–651 (2020), 10.1037/xlm0000966.
43. B. B. Murdock, The serial position effect of free recall. *J. Exp. Psychol.* **64**, 482–488 (1962).
44. E. Tulving, T. Y. Arbuckle, Sources of intratrial interference in immediate recall of paired associates. *J. Verbal Learn. Verbal Behav.* **1**, 321–334 (1963).
45. A. Aka, T. D. Phan, M. J. Kahana, Predicting recall of words and lists. *J. Exp. Psychol.: Learn. Mem. Cognit.* **47**, 765–784 (2020), 10.1037/xlm0000964.
46. Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, A. Oliva, Intrinsic and extrinsic effects on image memorability. *Vis. Res.* **116**, 165–178 (2015).
47. T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol.: Gen.* **139**, 558–578 (2010).
48. H. von Restorff, Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychol. Forsch.* **18**, 299–342 (1933).
49. U. Hasson, O. Furman, D. Clark, Y. Dudai, L. Davachi, Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron* **57**, 452–462 (2008).
50. J. Chen *et al*., Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* **20**, 115–125 (2017).
51. I. S. Jones, K. P. Kording, Quantifying the role of neurons for behavior is a mediation question. *Behav. Brain Sci.* **42**, e233 (2019).
52. T. J. VanderWeele, Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs (2021). eprint: 2007.00520.
53. H. M. Blalock, Making causal inferences for unmeasured variables from correlations among indicators. *Am. J. Sociol.* **69**, 53–62 (1963).
54. K. A. Bollen, Latent variables in psychology and the social sciences. *Ann. Rev. Psychol.* **53**, 605–634 (2002).
55. B. N. Sánchez, E. Budtz-Jørgensen, L. M. Ryan, H. Hu, Structural equation models: A review with applications to environmental epidemiology. *J. Am. Stat. Assoc.* **100**, 1443–1455 (2005).
56. R. Silva, R. Scheines, C. Glymour, P. Spirtes, "Learning measurement models for unobserved variables" in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, UAI 2003* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002), pp. 543–550. event-place: Acapulco, Mexico.
57. J. Rissman, A. D. Wagner, Distributed representations in memory: Insights from functional brain imaging. *Ann. Rev. Psychol.* **63**, 101–128 (2012).
58. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2** (2008), 10.3389/neuro.06.004.2008.
59. S. Edelman, K. Grill-Spector, T. Kushnir, R. Malach, Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology* **26**, 309–321 (1998).
60. G. Xue *et al*., Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding. *Cereb. Cortex* **23**, 1562–1571 (2013).
61. E. J. Ward, M. M. Chun, B. A. Kuhl, Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *J. Neurosci.* **33**, 14749–14757 (2013).
62. R. M. Visser, H. S. Scholte, T. Beemsterboer, M. Kindt, Neural pattern similarity predicts long-term fear memory. *Nat. Neurosci.* **16**, 388–390 (2013).
63. H. Bruett, R. C. Calloway, N. Tokowicz, M. N. Coutanche, Neural pattern similarity across concept exemplars predicts memory after a long delay. *NeuroImage* **219**, 117030 (2020).
64. R. N. van den Honert, G. McCarthy, M. K. Johnson, Reactivation during encoding supports the later discrimination of similar episodic memories: Reactivation Supports Mnemonic Discrimination. *Hippocampus* **26**, 1168–1178 (2016).
65. E. Cowan *et al*., Sleep spindles promote the restructuring of memory representations in ventromedial prefrontal cortex through enhanced hippocampal-cortical functional connectivity. *J. Neurosci.* **40**, 1909–1919 (2020).
66. T. Davis, G. Xue, B. C. Love, A. R. Preston, R. A. Poldrack, Global neural pattern similarity as a common basis for categorization and recognition memory. *J. Neurosci.* **34**, 7472–7484 (2014).
67. Y. Ezzyat, L. Davachi, Neural evidence for representational persistence within events. *J. Neurosci.* **41**, JN–RM–0073–21 (2021).
68. K. F. LaRocque *et al*., Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *J. Neurosci.* **33**, 5466–5474 (2013).
69. A. Tompary, L. Davachi, Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* **96**, 228–241.e5 (2017).
70. X. Xiao, Q. Dong, C. Chen, G. Xue, Neural pattern similarity underlies the mnemonic advantages for living words. *Cortex* **79**, 99–111 (2016).
71. G. E. Koch, J. P. Paulus, M. N. Coutanche, Neural patterns are more similar across individuals during successful memory encoding than during failed memory encoding. *Cereb. Cortex* **30**, 3872–3883 (2020).
72. T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
73. J. R. Anderson, S. Betts, J. L. Ferris, J. M. Fincham, Neural imaging to track mental states while using an intelligent tutoring system. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7018–7023 (2010).
74. T. O. Nelson, J. Dunlosky, Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory* **2**, 325–335 (1994).
75. P. J. Grimaldi, M. A. Pyc, K. A. Rawson, Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian-English paired associates. *Behav. Res. Methods* **42**, 634–642 (2010).
76. T. Y. Arbuckle, L. L. Cuddy, Discrimination of item strength at time of presentation. *J. Exp. Psychol.* **81**, 126–131 (1969).
77. E. A. Lovelace, Metamemory: Monitoring future recallability during study. *J. Exp. Psychol.: Learn. Mem. Cognit.* **10**, 756–766 (1984).
78. T. O. Nelson, J. Dunlosky, When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "Delayed-JOL Effect". *Psychol. Sci.* **2**, 267–271 (1991).
79. R. J. Mislevy, Exploiting auxiliary infornlation about examinees in the estimation of item parameters. *Appl. Psychol. Meas.* **11**, 81–91 (1987).
80. W. Van den Noortgate, P. De Boeck, Assessing and explaining differential item functioning using logistic mixed models. *J. Educ. Behav. Stat.* **30**, 443–464 (2005).
81. T. A. Cleary, T. L. Hilton, An investigation of item bias. *Educ. Psychol. Meas.* **28**, 61–75 (1968).
82. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, vol. 1. 1966).
83. J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
84. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
85. A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance" in *Twenty-First International Conference on Machine Learning - ICML 2004 (ACM Press, Banff, Alberta, Canada, 2004), p. 78.*
86. S. L. Cessie, J. C. V. Houwelingen, Ridge estimators in logistic regression. *Appl. Stat.* **41**, 191 (1992).
87. B. A. Kuhl, J. Rissman, A. D. Wagner, Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia* **50**, 458–469 (2012).
88. J. Rissman, H. T. Greely, A. D. Wagner, Detecting individual memories through the neural decoding of memory states and past experience. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9849–9854 (2010).
89. S. Chakravarty, Y. Y. Chen, J. B. Caplan, Predicting memory from study-related brain activity. *J. Neurophysiol.* **124**, jn.00193.2020 (2020).
90. P. T. Reiss, L. Huo, Y. Zhao, C. Kelly, R. T. Ogden, Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Ann. Appl. Stat.* **9**, 1076–1101 (2015).
91. D. M. Potter, A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat. Med.* **24**, 693–708 (2005).
92. J. A. Mumford, B. O. Turner, F. G. Ashby, R. A. Poldrack, Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
93. C. M. Theobald, Generalizations of mean square error applied to ridge regression. *J. R. Stat. Soc.: Ser. B (Methodol.)* **36**, 103–106 (1974).
94. N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3863–3868 (2006).
95. Q. Wang, B. Cagna, T. Chaminade, S. Takerkart, Inter-subject pattern analysis: A straightforward and powerful scheme for group-level MVPA. *NeuroImage* **204**, 116205 (2020).
96. J. V. Haxby, J. S. Guntupalli, S. A. Nastase, M. Feilong, Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
97. R. M. Dawes, The robust beauty of improper linear models in decision making. *Am. Psychol.* **34**, 571–582 (1979).
98. F. L. Schmidt, The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educ. Psychol. Meas.* **31**, 699–714 (1971).

99. A. Schaefer *et al*., Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
100. W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).
101. D. Marr, Simple memory: A theory for archicortex. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **262**, 23–81 (1971).
102. K. A. Norman, R. C. O'Reilly, Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* **110**, 611–646 (2003).
103. L. Litman, L. Davachi, Distributed learning enhances relational memory consolidation. *Learn. Mem.* **15**, 711–716 (2008).
104. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
105. M. Stone, Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc.: Ser. B (Methodol.)* **36**, 111–133 (1974).
106. T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
107. F. Pereira, M. Botvinick, Information mapping with pattern classifiers: A comparative study. *NeuroImage* **56**, 476–496 (2011).
108. E. J. G. Pitman, Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **4**, 119 (1937).
109. J. Stelzer, Y. Chen, R. Turner, Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage* **65**, 69–82 (2013).
110. P. Golland, B. Fischl, "Permutation tests for classification: Towards statistical significance in image-based studies" in *Information Processing in Medical Imaging. Lecture Notes in Computer Science*, G. Goos, J. Hartmanis, J. van Leeuwen, C. Taylor, J. A. Noble, Eds. (Springer, Heidelberg, 2003), vol. 2732, pp. 330–341.
111. B. A. Spellman, R. A. Bjork, When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychol. Sci.* **3**, 315–317 (1992).
112. W. L. Kelemen, C. A. Weaver, Enhanced memory at delays: Why do judgments of learning improve over time? *J. Exp. Psychol.: Learn. Mem. Cognit.* **23**, 1394–1409 (1997).
113. M. G. Rhodes, S. K. Tauber, The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychol. Bull.* **137**, 131–148 (2011).
114. S. K. U. Tauber, J. Dunlosky, K. A. Rawson, The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Exp. Psychol.* **62**, 254–263 (2015).
115. J. Haxby *et al*., A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
116. M. Blanco-Pozo, T. Akam, M. Walton, Dopamine reports reward prediction errors, but does not update policy, during inference-guided choice. *Neuroscience* (2021, preprint).
117. O. Esteban *et al*., fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
118. K. Gorgolewski *et al*., Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Front. Neuroinf.* **5** (2011), 10.3389/fninf.2011.00013.
119. N. J. Tustison *et al*., N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
120. M. Reuter, H. D. Rosas, B. Fischl, Highly accurate inverse consistent registration: A robust approach. *NeuroImage* **53**, 1181–1196 (2010).
121. A. M. Dale, B. Fischl, M. I. Sereno, Cortical surface-based analysis. *NeuroImage* **9**, 179–194 (1999).
122. A. Klein *et al*., Mindboggling morphometry of human brains. *PLOS Comput. Biol.* **13**, e1005350 (2017).
123. V. Fonov *et al*., Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage* **54**, 313–327 (2011).
124. B. B. Avants *et al*., A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* **54**, 2033–2044 (2011).
125. Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
126. H. Cox, Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 8 (1997).
127. D. N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
128. M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**, 825–841 (2002).
129. J. D. Power *et al*., Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
130. Y. Behzadi, K. Restom, J. Liau, T. T. Liu, A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
131. C. Lanczos, A precision approximation of the gamma function. *J. Soc. Ind. Appl. Math. Ser. B Numer. Anal.* **1**, 86–96 (1964).
132. M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
133. C. B. Martin, D. Douglas, R. N. Newsome, L. L. Man, M. D. Barense, Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife* **7**, e31873 (2018).
134. L. Chen *et al*., The visual word form area (VWFA) is part of both language and attention circuitry. *Nat. Commun.* **10**, 5601 (2019).
135. M. Ritchey, M. E. Montchal, A. P. Yonelinas, C. Ranganath, Delay-dependent contributions of medial temporal lobe regions to episodic memory retrieval. *eLife* **4**, e05025 (2015).
136. K. J. Friston *et al*., Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).
137. M. A. Lindquist, J. Meng Loh, L. Y. Atlas, T. D. Wager, Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* **45**, S187–S198 (2009).
138. W. McKinney, Data Structures for Statistical Computing in Python. (Austin, Texas) (2010), pp. 56–61.
139. S. van der Walt, S. C. Colbert, G. Varoquaux, The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
140. H. Wickham *et al*., Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
141. M. Kuhn, H. Wickham, Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles (2020).
142. S. Bobadilla-Suarez, C. Ahlheim, A. Mehrotra, A. Panos, B. C. Love, Measures of neural similarity. *Comput. Brain Behav.* **3** (2019).
143. A. Walther *et al*., Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200 (2016).
144. R. A. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507 (1915).
145. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015), 10.18637/jss.v067.i01.
146. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010), 10.18637/jss.v033.i01.