Report

Current Biology

Cross-Situational Learning Is Supported by Propose-but-Verify Hypothesis Testing

Highlights

- Subjects learned word-object associations across multiple exposures during fMRI
- Learning involved regions associated with working memory and reward processing
- RSA showed that learning was mediated by a propose-but-verify mechanism

Authors

Sam C. Berens, Jessica S. Horst,
Chris M. Bird

Correspondence

sam.berens@york.ac.uk (S.C.B.),
chris.bird@sussex.ac.uk (C.M.B.)

In Brief

Using model-based representation similarity analyses of fMRI data, Berens et al. find evidence for cross-situational word learning mediated by a propose-but-verify mechanism in the hippocampus. This suggests that adults rely on their episodic memory to track a limited number of associations when learning new words across events.

Berens et al., 2018, Current Biology *28*, 1–5
April 2, 2018 © 2018 The Author(s). Published by Elsevier Ltd.
https://doi.org/10.1016/j.cub.2018.02.042

Current Biology
Report

Cross-Situational Learning Is Supported by Propose-but-Verify Hypothesis Testing

Sam C. Berens,[1,2,3,*] Jessica S. Horst,[1] and Chris M. Bird[1,*]
[1]School of Psychology, University of Sussex, Falmer BN1 9QH, UK
[2]Department of Psychology, University of York, York YO10 5DD, UK
[3]Lead contact
*Correspondence: sam.berens@york.ac.uk (S.C.B.), chris.bird@sussex.ac.uk (C.M.B.)

https://doi.org/10.1016/j.cub.2018.02.042

SUMMARY

When we encounter a new word, there are often multiple objects that the word might refer to [1]. Nonetheless, because names for concrete nouns are constant, we are able to learn them across successive encounters [2, 3]. This form of "cross-situational" learning may result from either associative mechanisms that gradually accumulate evidence for each word-object association [4, 5] or rapid propose-but-verify (PbV) mechanisms where only one hypothesized referent is stored for each word, which is either subsequently verified or rejected [6, 7]. Using model-based representation similarity analyses of fMRI data acquired during learning, we find evidence for learning mediated by a PbV mechanism. This learning may be underpinned by rapid pattern-separation processes in the hippocampus. Our findings shed light on the psychological and neural processes that support word learning, suggesting that adults rely on their episodic memory to track a limited number of word-object associations.

RESULTS

Humans have a huge capacity for learning new information. Remarkably, such learning can occur even when information is incompletely provided. For example, when encountering an unfamiliar word, there is often an almost limitless number of objects or concepts that it could hypothetically refer to [1]. Given this, learning the meaning of a word from a single instance is impossible. Nonetheless, after repeated exposures across different situations, both adults and children are able to learn word-referent associations [2, 3]. Although learning names for abstract words is likely more complicated, cross-situational learning is thought to underpin the learning of name-object associations of concrete nouns during early childhood [4]. Here, we sought to understand the mechanisms supporting it.

We scanned adult participants as they performed a cross-situational learning task involving 9 novel associations between obscure objects and pseudowords. On each learning event, participants saw three unfamiliar objects and heard their corresponding pseudowords (Figure 1A). There was no relationship between the location of the objects and word order. Therefore, to learn the associations, information had to be carried over trials. Learning events were grouped into 6 blocks and each of these was followed by a set of 9-alternative forced-choice (9-AFC) test trials to assess whether correct associations had been learned (Figure 1B). To control for the visual and motor aspects of the task, a separate set of 9 word-object pairs were pre-learned before scanning and presented/tested in the same way as the to-be-learned words and objects. For a full description of the task, see STAR Methods.

Model-free Analysis of Learning-Related Brain Activity

Participants learned the word-object associations across the 6 learning blocks (Figures 1C and 1D), consistent with previous findings (e.g., [3]). In the first fMRI analysis, we wished to identify the brain regions involved in cross-situational learning. In particular, we were interested in regions that were most active during trials when most learning occurred. Thus, we conducted a model-free analysis of the fMRI data to identify correlations between blood-oxygen-level dependent (BOLD) activity and the amount of learning taking place. Specifically, the amount of learning that occurred during each learning block (i.e., the learning rate) was taken as the change in 9-AFC test trial performance that occurred between blocks (see Figure 1D). This measure was then correlated with BOLD activity from the learning events.

Activity was positively correlated with learning in a number of frontal and parietal regions as well as the dorsal striatum bilaterally and fusiform gyrus (Figures 1E and 1F; Table S1; statistical image available at https://neurovault.org/collections/3002/). These regions are functionally connected during rest (see Figure S1) and are co-activated during many fMRI tasks involving effortful processes (e.g., [8, 9]). Importantly, the regions that we identify are commonly recruited when learning associations between items (e.g., [10, 11]). These findings suggest that on a general level, as detected by regional changes in activity, adult cross-situational learning draws on a similar set of processes as other tasks involving explicit learning of associations, such as working memory, attention, and reward processing.

Activation Patterns Centered on the Hippocampus Support PbV Learning

Next, we carried out a model-based analysis of the fMRI data to test two competing models of cross-situational learning. According to "global" or "associative models," all co-occurrences of words and objects are maintained and updated across each encounter in the form of weighted connections between words

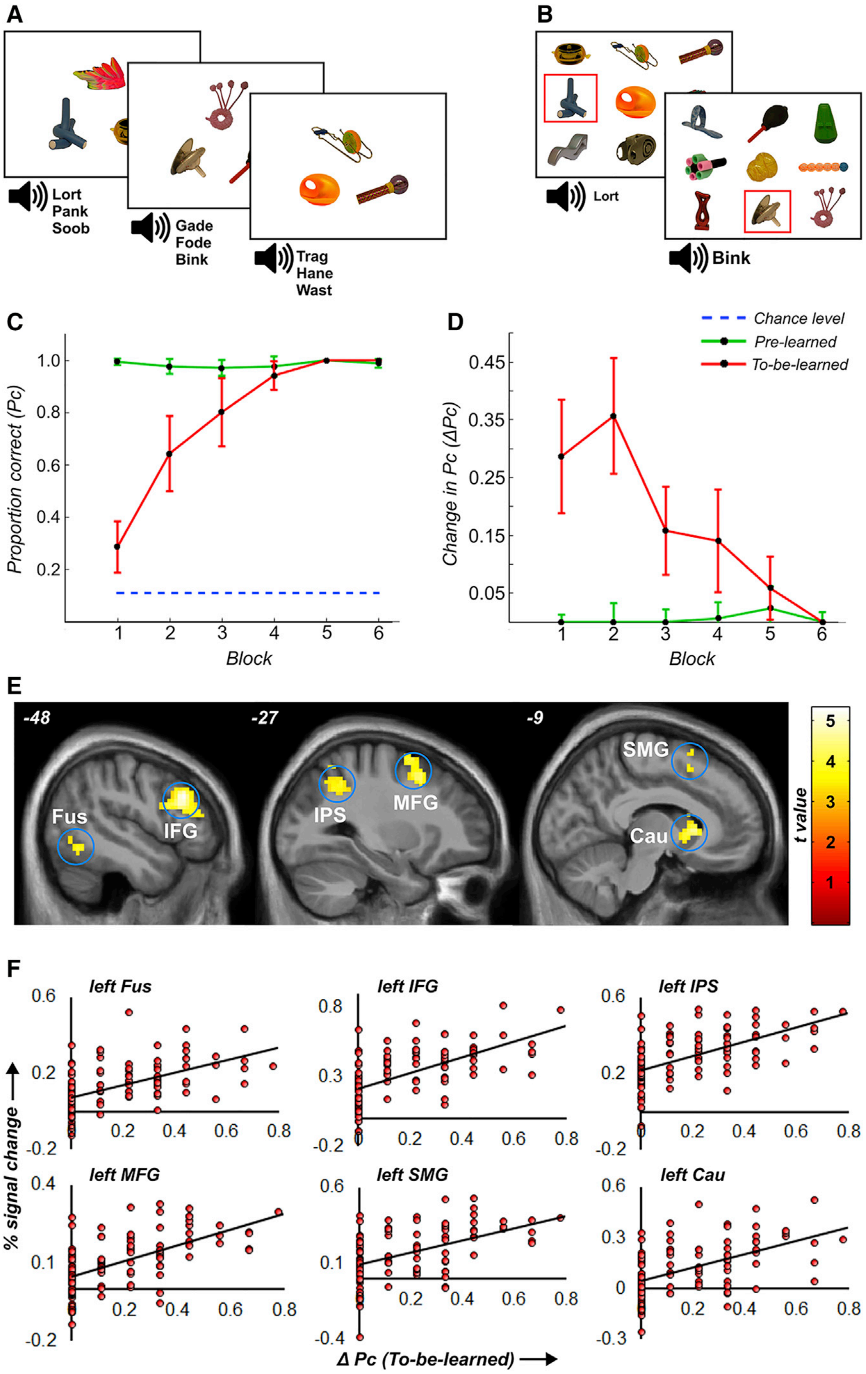

Figure 1. Details of In-Scanner Task and Model-free Analyses

(A) On each learning trial, 3 unfamiliar objects were presented on screen, and their corresponding pseudowords were presented auditorily. There was no relationship between the object locations on screen and word order. Therefore, the word-object associations had to be tracked across multiple ambiguous learning events.

(B) After each learning block, all of the to-be-learned associations, as well as the pre-learned control associations, were assessed via 9-AFC trials.

(C) Participants gradually learned the experimental word-object pairs over the 6 blocks. Performance on the pre-learned pairs was at ceiling.

(D) Amount learned during the block, indexed by the changes in performance from the beginning to the end of each block. Most learning took place in the first 2 blocks. Error bars indicate 95% confidence intervals.

(E) Regions where BOLD activity correlated with learning during the encoding events. We observed effects in a fronto-parietal network as well as the fusiform gyrus and the head of the caudate nucleus.

(F) Plots displaying % signal change estimates during the learning trials and change in performance statistics for the regions identified in (E).

and objects (e.g., [4, 5]). As such, early in the course of learning, some associations will exist between each word and every object it has been seen with. However, after repeated encounters of the words and objects, the correct associations will emerge and dominate. A feature of associative models is that some information is learnt on each encounter and that if a dominant association proves to be incorrect, other plausible associations will be available, allowing the correct association to be strengthened over time. In contrast, "local models" posit that only a very limited number of hypotheses about the word-object associations are carried forward until they are verified or disconfirmed on later encounters (e.g., [6]). The strongest version of a local model is propose-but-verify (PbV) hypothesis testing (see [7]). Under such a model, when an individual hears a word, she arbitrarily proposes an association between the word and one of the objects and then verifies or rejects this proposal on future encounters. Associations between the word and other possible referent objects are not stored. Therefore, under PbV learning, acquisition of the word-object associations occurs in an all-or-nothing manner such that associations are either correctly guessed and verified or not learnt at all.

Critically, these theories make opposing predictions about the neural representations of word-object associations during learning. Specifically, they differ in their predictions about when representations of the correct word-object associations are created. Under any learning model, once a word has been associated with a particular object, it acquires a unique meaning, whereas unfamiliar words remain meaningless. Thus, after learning, the representation of a word will be distinctly different from all others. This process, whereby distinct memory representations are created from similar learning events, is known as pattern separation [12, 13]. At a neural level, pattern separation involves the hippocampus [12, 13]. Interestingly, the hippocampus is implicated in both rapid and

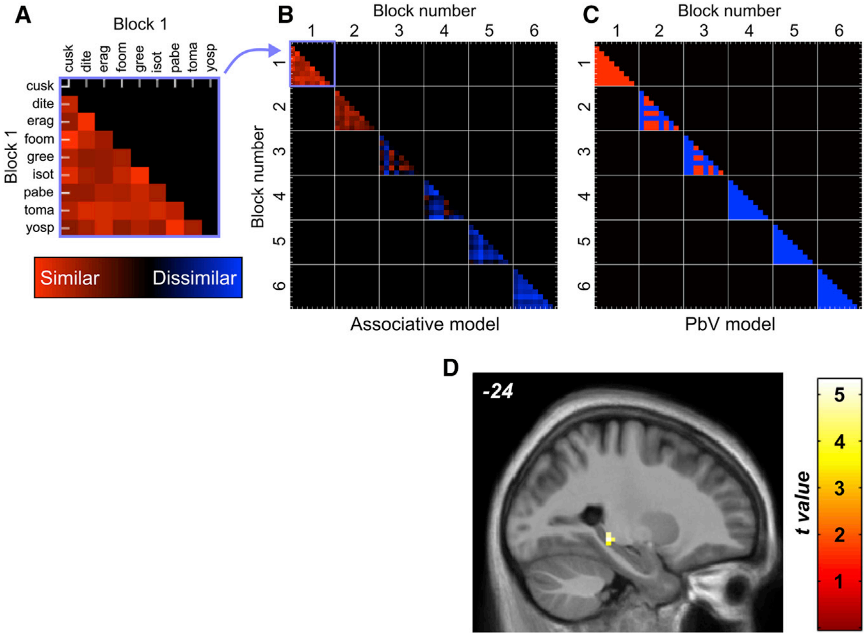

Figure 2. Model-Based Representational Similarity Analyses of Test Trials

(A) Example RSA contrast matrix showing predicted similarity for every possible pair of the 9 objects and 9 to-be-learned pseudowords during the first run.

(B) Example RSA contrast matrix according to the associative model across all 6 test blocks. Unlearnt associations are equally similar to each other (red), and learned associations are equally dissimilar (blue). Each participant's behavioral data were used to construct their own matrix, which predicts the similarity in patterns of BOLD activity. Under this model, the transition from similar to dissimilar representations proceeds gradually.

(C) Example RSA contrast matrix according to the PbV model across all 6 test blocks. Under PbV, learning is assumed to occur within the learning block prior to making a correct response, resulting in an abrupt representational switch between unlearnt and learned pairs.

(D) Searchlights centered on a region of the left hippocampus showed changes in representational similarity that are consistent with predictions of the PbV learning model (peak voxel MNI: −24, −33, −6). T values indicate the size of this effect across participants (thresholded at $p < 0.001$). No above-threshold effects were identified for the associative learning model. Unthresholded statistical images are available at https://neurovault.org/collections/3002/.

gradual pattern separation [14–16] as well as in associative or relational memory more generally (e.g., [17]). Therefore, it is possible that the hippocampus may play a role in cross-situational learning via associative and/or PbV mechanisms.

PbV models predict that learning is rapid; initially, an association between a word and object is arbitrarily chosen, and if verified on subsequent encounters, it is retained. This will cause the neural representation of the word to rapidly change from being similar to being dissimilar from the other words. In contrast, associative models predict that word-object associations emerge gradually, with evidence stored about all word and object co-occurrences (although, we note that under some models, attentional biases may accelerate acquisition at particular time points, e.g.[5]). Accordingly, the neural representations should become dissimilar gradually over successive learning blocks.

To test these predictions, patterns of fMRI BOLD activity for each pseudoword during each test block were obtained, and pairwise correlations of these patterns were computed throughout the brain. The correlations were then compared to the predicted similarity of every word pair according to the two models, using representational similarity analysis [18]. Figures 2B and 2C illustrate the predictions of the associative and PbV models for a representative participant. Under the associative model, the correct associations emerge gradually, being strengthened with each encounter of the word-object pair. The associative model also predicts that incorrect word-object associations will be made if they happen to co-occur on several trials by chance. By contrast, the PbV model predicts that the representations rapidly change from being equally similar to all others before they have been learnt to being dissimilar after learning. Importantly, the predictions of each model reflect the timecourse at which changes in representational similarity should occur rather than the absolute level of representational change that may be expected. As such, the only source of predictive power afforded to either model stemmed from stipulating this timecourse as accurately as a possible. (For more details about how the data were modeled and analyzed, see STAR Methods.)

We found evidence for the PbV model but not the associative model. A whole-brain searchlight representational similarity analysis (RSA) revealed a significant fit to the predictions of the PbV model within our *a priori* region of interest (ROI) in the hippocampus (see Figure 2C; $t_{18} = 5.36$, $p = 0.013$, corrected for family-wise error (FWE) within a bilateral hippocampal ROI). Furthermore, goodness-of-fit statistics indicated stronger support for the PbV model over the associative model within this region ($\Delta BIC = 7.92$; [19]). Due to its size and shape, searchlights where the centers fall within the hippocampus additionally include adjacent extrahippocampal voxels. When we excluded all extrahippocampal voxels from the region showing the effect reported above, the goodness-of-fit statistics still indicated very strong support for the PbV model over the associative model ($\Delta BIC = 23.61$; based on an average of 32.2 voxels), although the fit to the PbV model within this restricted region was only significant at an uncorrected threshold ($t_{18} = 2.99$, $p = 0.001$). To assess the contribution of extrahippocampal voxels to the fit to the PbV model, we reran the whole-brain analysis while excluding all voxels within the hippocampal ROI. This analysis revealed no significant effects anywhere in the brain, even at lenient threshold of $p < 0.002$.

To test whether there was evidence for either the PbV or associative models across the group of brain regions that exhibited learning-related effects in the model-free analysis above, we examined representational similarity across all voxels within each region; no significant effects were found in any regions.

Thus, our model-based fMRI analysis provided evidence that cross-situational learning is supported by a PbV mechanism. Representations of the words in a region focused on the left posterior hippocampus rapidly became dissimilar when the association between the word and object had been learned. This is consistent with a verification model that may be underpinned by hippocampally mediated pattern separation. During debriefing, 10 of 19 participants reported using a hypothesis-testing or process-of-elimination strategy, while the remaining participants did not report using any particular strategy. Overall, our results suggest that when learning associations across situations, adults store a limited number of hypotheses about word-object associations that are verified or rejected on later encounters.

DISCUSSION

There are a number of advantages to a PbV "local" style of learning compared to one that monitors all co-occurrences of items. PbV models are better placed to exploit rare but informative learning events, while models that keep track of all options dilute such instances [20]. There are also computational advantages to a learning system that actively chooses to store a limited amount of information that is either verified or rejected in the future. The alternative—to keep a running tally of the number of instances each object occurred with each referent—requires a large capacity and ultimately results in the storage of much redundant information.

Our findings parallel recent discoveries in the decision-making literature. Classic theories suggest that the outcomes of previous decisions result in the accumulation of evidence for the value of different options [21]. When making a decision, the evidence for and against a range of choices can be compared in order to make the correct decision. However, several researchers have questioned the existence of such a knowledge base and argued that sampling a limited number of recent events gives sufficient information on which to base a decision (e.g., [22, 23]). Retrieving individual episodic memories is also the preferred basis of decision-making in some situations [24]. This is similar to a PbV learning mechanism, which relies on episodic memory for a limited number of hypothesized associations.

Recently, a number of authors have advocated less strong versions of both PbV and associative models [20, 25, 26]. For example, Stevens et al. [20] proposed a learning mechanism that tracks more than one hypothesized word-object pairing but pursues the highest-valued pairing at the expense of others. By contrast, Yurovsky and Frank [25] argued that learning mechanisms are limited by psychological constraints on memory and attention; PbV is favored when there are many associations, but when there are few, a more associative style of learning predominates. Lastly, Kachergis and colleagues [5, 26] found that associative models that bias attention toward both novel and familiar word-object pairings (rather than record all co-occurrences indiscriminately) fare better than PbV models at predicting learning trajectories. It is not yet clear whether both PbV and associative mechanisms can operate separately in parallel, with greater weight placed on one or the other mechanism according to task demands, or whether there is a unitary learning mechanism that combines elements of both. Functional imaging offers an opportunity to resolve such issues (see [27] for an analogous example in the spatial memory domain).

The model-based fMRI effect was centered on the left posterior hippocampus, although the searchlight included immediately adjacent areas. According to the complimentary learning systems (CLS) theory, the hippocampus is able to support rapid learning of similar materials via pattern separation [28, 29]. Furthermore, neuropsychological evidence points toward a necessary role for the hippocampus in the acquisition of new semantic knowledge, including vocabulary acquisition, which is consistent with our findings [30–32]. Pattern separation enables the creation of non-overlapping representations of memories that are otherwise highly similar. A recent version of this theory identified roles for the hippocampus in both rapid and gradual pattern separation [16]. In our study, we only found evidence for the rapid creation of novel memory representations, which was predicted by the PbV model. Nevertheless, this does not rule out the possibility that more gradual associative learning was also taking place, supported either by processing in the hippocampus or elsewhere in the brain.

In sum, we present fMRI evidence that cross-situational learning is supported by PbV mechanisms that are underpinned by hippocampal processing. More broadly, remembering a limited sample of hypotheses about possible associations, and testing these hypotheses against future encounters, may be a core method of acquiring new declarative knowledge, particularly when learning information you know you will be tested on. This is because it capitalizes on humans' well-developed episodic memory system for remembering individual events.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Materials
  - Procedure
  - MRI Acquisition
  - Image Pre-processing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Behavioral data
  - Univariate imaging analyses
  - Representational similarity analysis
  - Propose-but-verify model
  - Associative learning model
  - Imaging thresholds
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes one table and one figure and can be found with this article online at https://doi.org/10.1016/j.cub.2018.02.042.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (337822 to C.M.B.) and an Economic and Social Research Council studentship (ES/J500173/1 to S.C.B.).

AUTHOR CONTRIBUTIONS

S.C.B., J.S.H., and C.M.B. designed the experiment. S.C.B. coded the experimental task, performed the experiment, and analyzed the data. All authors discussed the results and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 18, 2017
Revised: January 16, 2018
Accepted: February 16, 2018
Published: March 15, 2018

REFERENCES

1. Quine, W.V. (1960). Word and object (MIT Press).
2. Smith, L., and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition *106*, 1558–1568.
3. Yu, C., and Smith, L.B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychol. Sci. *18*, 414–420.
4. McMurray, B., Horst, J.S., and Samuelson, L.K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychol. Rev. *119*, 831–877.
5. Kachergis, G., Yu, C., and Shiffrin, R.M. (2012). An associative model of adaptive inference for learning word-referent mappings. Psychon. Bull. Rev. *19*, 317–324.
6. Medina, T.N., Snedeker, J., Trueswell, J.C., and Gleitman, L.R. (2011). How words can and cannot be learned by observation. Proc. Natl. Acad. Sci. USA *108*, 9014–9019.
7. Trueswell, J.C., Medina, T.N., Hafri, A., and Gleitman, L.R. (2013). Propose but verify: fast mapping meets cross-situational word learning. Cognit. Psychol. *66*, 126–156.
8. Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., and Eickhoff, S.B. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. Neuroimage *60*, 830–846.
9. Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., and Raichle, M.E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. USA *102*, 9673–9678.
10. Anderson, J.R., Byrne, D., Fincham, J.M., and Gunn, P. (2008). Role of prefrontal and parietal cortices in associative learning. Cereb. Cortex *18*, 904–914.
11. Corlett, P.R., Aitken, M.R., Dickinson, A., Shanks, D.R., Honey, G.D., Honey, R.A., Robbins, T.W., Bullmore, E.T., and Fletcher, P.C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. Neuron *44*, 877–888.
12. Rolls, E.T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. Front. Syst. Neurosci. *7*, 74.
13. Yassa, M.A., and Stark, C.E. (2011). Pattern separation in the hippocampus. Trends Neurosci. *34*, 515–525.
14. Bakker, A., Kirwan, C.B., Miller, M., and Stark, C.E. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. Science *319*, 1640–1642.
15. Milivojevic, B., Varadinov, M., Vicente Grabovetsky, A., Collin, S.H., and Doeller, C.F. (2016). Coding of Event Nodes and Narrative Context in the Hippocampus. J. Neurosci. *36*, 12412–12424.
16. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049.
17. Eichenbaum, H., and Cohen, N.J. (2001). From conditioning to conscious recollection: Memory systems of the brain (Oxford University Press).
18. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.
19. Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. J. Am. Stat. Assoc. *90*, 773–795.
20. Stevens, J.S., Gleitman, L.R., Trueswell, J.C., and Yang, C. (2017). The Pursuit of Word Meanings. Cogn. Sci. *41* (*Suppl 4*), 638–676.
21. von Neumann, J., and Morgenstern, O. (1944). Theory of Games and Economic Behaviour (Princeton University Press).
22. Giguère, G., and Love, B.C. (2013). Limits in decision making arise from limits in memory retrieval. Proc. Natl. Acad. Sci. USA *110*, 7613–7618.
23. Stewart, N., Chater, N., and Brown, G.D. (2006). Decision by sampling. Cognit. Psychol. *53*, 1–26.
24. Bornstein, A.M., Khaw, M.W., Shohamy, D., and Daw, N.D. (2017). Reminders of past choices bias decisions for reward in humans. Nat. Commun. *8*, 15958.
25. Yurovsky, D., and Frank, M.C. (2015). An integrative account of constraints on cross-situational learning. Cognition *145*, 53–62.
26. Kachergis, G., Yu, C., and Shiffrin, R.M. (2017). A bootstrapping model of frequency and context effects in word learning. Cogn. Sci. *41*, 590–622.
27. Burgess, N. (2006). Spatial memory: how egocentric and allocentric combine. Trends Cogn. Sci. *10*, 551–557.
28. Kumaran, D., and McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. *119*, 573–616.
29. McClelland, J.L., McNaughton, B.L., and O’Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419–457.
30. Bayley, P.J., O’Reilly, R.C., Curran, T., and Squire, L.R. (2008). New semantic learning in patients with large medial temporal lobe lesions. Hippocampus *18*, 575–583.
31. Kitchener, E.G., Hodges, J.R., and McCarthy, R. (1998). Acquisition of post-morbid vocabulary and semantic facts in the absence of episodic memory. Brain *121*, 1313–1327.
32. O’Kane, G., Kensinger, E.A., and Corkin, S. (2004). Evidence for semantic learning in profound amnesia: an investigation with patient H.M. Hippocampus *14*, 417–425.
33. Horst, J.S. (2009). The Novel Object and Unusual Name (NOUN) database. http://www.sussex.ac.uk/wordlab/noun.
34. Andersson, J.L., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage *20*, 870–888.
35. Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. Neuroimage *16*, 217–240.
36. Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. Neuroimage *38*, 95–113.
37. Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., and Kiebel, S. (2005). Mixed-effects and fMRI studies. Neuroimage *24*, 244–252.
38. Oosterhof, N.N., Connolly, A.C., and Haxby, J.V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front. Neuroinform. *10*, 27.
39. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage *15*, 273–289.

Please cite this article in press as: Berens et al., Cross-Situational Learning Is Supported by Propose-but-Verify Hypothesis Testing, Current Biology (2018), https://doi.org/10.1016/j.cub.2018.02.042

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Unthresholded statistical maps for all analyses	This paper	https://neurovault.org/collections/3002/
Software and Algorithms		
MATLAB	MathWorks	https://www.mathworks.com
Statistical Parametric Mapping 8	FIL (UCL)	www.fil.ion.ucl.ac.uk/spm
CoSMoMVPA	[38]	http://www.cosmomvpa.org/
Other		
NOUN Database	[33]	http://www.sussex.ac.uk/wordlab/noun
Automated Anatomical Labeling 2	[39]	http://www.gin.cnrs.fr/en/tools/aal-aal2/
1.5T Siemens Avanto scanner	Siemens	https://www.siemens.com/global/em/home.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Sam Berens (sam.berens@york.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Twenty-three right-handed, native English-speaking students were recruited from the University of Sussex. All gave written informed consent and were reimbursed for their time. Participants had either normal or corrected-to-normal vision and reported no history of neurological or psychiatric illness. Data from four participants could not be included in the final sample due to problems with fMRI data acquisition (1 participant), and a failure to learn more than 3 of 9 word-object pairs during the in-scanner task (3 participants). These latter participants were excluded since their level of performance could significantly rule out the possibility that they were responding randomly ($p = 0.069$, $n = 9$, $k = 3$, $E[k] = 1$). Additionally, their performance was well below that of all other participants who reached ceiling by the penultimate test block. As such, analyses included data from 19 participants (11 males) with a mean age of 25.4 years (SD = 4.0). The study was approved by the Brighton and Sussex Medical School's Research Governance and Ethics Committee.

METHOD DETAILS

Materials

Stimuli were 18 color photographs of obscure objects (e.g., rocket air blower) and 18 four-letter pseudowords (e.g., "Ospi") selected from the NOUN Database [33]. Prior to each session, these stimuli were randomly paired to form 18 word-object associations. Each pair was then allocated to one of two groups; a "pre-learned" set and a "to-be-learned" set (9 pairs in each). All photographs had a resolution of 240 × 240 pixels (in-scanner visual angle: $\sim 8^\circ$) and were taken against a white background. Pseudowords were presented auditorily via headphones and spoken by a neutral female voice (equated for perceived loudness).

Procedure

Pre-scanner training

We wished to include a control task, consisting of word-object pairs that had been pre-learned before scanning (a "no learning" control that was otherwise identical to the cross-situation learning task). To match the two tasks, the pseudowords and objects used in both tasks needed to be equally familiar – otherwise task differences might be caused by differing responses to the novelty/familiarity of the stimuli. Although this introduces a difference between our study and other cross-situation learning experiments, we do not think that it substantially affected performance on the task, since accuracy was only a little above chance after the first training block of the in-scanner task.

Prior to scanning, word-object associations for pre-learned pairs were trained using an explicit encoding paradigm. Following a 2 s inter-trial interval, a single object was presented centrally for 6 s and during this time the corresponding pseudoword was heard. There were 5 such study events for each association (i.e., 45 in total) and these progressed in a random order. Subsequently, participants were tested on each association with a 9-alternative forced-choice (9-AFC) test trial. After being cued with a single pre-learned pseudoword, a 3x3 grid of all the pre-learned objects was displayed (as in Figure 1B). Participants used a cursor controlled via computer keyboard to select the target object.

To equate the level of familiarity between pre-learned and to-be-learned stimuli, participants also engaged in a familiarisation phase for the to-be-learned words and objects. This took the form of a recognition memory test similar to the explicit encoding procedure described above with the key difference being that each trial only presented either a word or an object but not both simultaneously. There were 5 study events for each of the 18 to-be-learned stimuli (i.e., 90 in total) and these progressed in a random order. A two-alternative forced-choice recognition test then followed where a single to-be-learned stimulus (the target) was presented alongside a same-modality, unstudied lure. During pseudoword test trials, the target and lure words (also taken from the NOUN Database) were presented before the text strings "First" and "Last" were displayed on screen. Participants then indicated which word was the target. During object test trials, target and lure objects were themselves displayed simultaneously and participants selected the target using a cursor. There was a single test trial for each of the 18 to-be-learned stimuli and these were sequenced at random. The order in which the pre-scanner tasks were run was counterbalanced between participants.

In-scanner task

The in-scanner task consisted of 6 blocks of learning events and 6 blocks of test trials, with each learning block followed by a test block. Individual learning events lasted for 6 s. During this time, 3 word-object pairs randomly sampled from either the pre-learned or to-be-learned stimulus sets were presented (Figure 1A). The objects were positioned randomly in one of 3 on-screen locations. Their corresponding pseudowords were presented in a random order. There was no indication of which object went with which word. Trials were separated by a variable (uniformly distributed) inter-trial interval of 3 - 7 s. Both pre-learned and to-be-learned trials occurred in a random (intermixed) order and were constructed so that no association was presented consecutively (as in [5]). Within learning blocks, each word-object association was repeated 3 times. As such, there were 18 learning events per block.

Test blocks consisted of 18 trials, one for each of the pre-learned and to-be-learned associations. Individual trials occurred as the 9-AFC test trials run during pre-scanner training; a 3x3 grid of all the to-be-learned (or pre-learned) objects was displayed. After being cued with a single pseudoword, a randomly positioned cursor appeared around an object (Figure 1B). After a 1100ms delay, participants could move the cursor around the grid and select the cued object with an MRI compatible button box. All trials occurred in a random order and were spaced with a variable (uniformly distributed) inter-trial interval of 2 - 4 s. Learning and test blocks were separated from one another with an inter-block interval of 6 s.

MRI Acquisition

All images were acquired on a 1.5 Tesla Siemens Avanto scanner equipped with a 32-channel phased array head coil. T2*-weighted scans were acquired with echo-planar imaging (EPI), 34 axial slices (approximately 30° to AC-PC line; interleaved) and the following parameters; repetition time = 2520 ms, echo time = 43 ms, flip angle = 90°, slice thickness = 3.6 mm, in-plane resolution = 3 × 3 mm. To allow for T1 equilibrium, the first 5 EPI volumes were acquired prior to the task starting and then discarded. Subsequently, a field map was captured to allow the correction of geometric distortions caused by field inhomogeneity (see the Image Pre-processing section below). Finally, for purposes of co-registration and image normalization, a whole-brain T1-weighted structural scan was acquired with a 1mm^3 resolution using a magnetization-prepared rapid gradient echo pulse sequence.

Image Pre-processing

Image pre-processing was performed in SPM8 (www.fil.ion.ucl.ac.uk/spm) and using custom written code in MATLAB (Mathworks). First, each subject's EPI volumes were corrected for inter-slice acquisition delay and spatially realigned to the first image in the time series. At the same time, images were corrected for field inhomogeneity based geometric distortions (as well as the interaction between motion and such distortions) using the *Realign and Unwarp* algorithms in SPM [34, 35]. For the analyses of univariate BOLD activations, EPI time series data were warped to MNI space using transformation parameters derived from the structural scans (with the DARTEL toolbox; [36]). Subsequently, the EPI volumes were spatially smoothed with an isotropic 8 mm FWHM Gaussian kernel prior to GLM analysis. For the representational similarity analysis, multivariate BOLD patterns of interest were estimates as the *t*-statistics resulting from a GLM of the unsmoothed EPI data in native space (see below).

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral data

Behavioral outputs from the in-scanner task were binary (correct versus incorrect) statistics relating to accuracy on each of the 9-AFC test trials (coded as 1 and 0, respectively). The proportion correct (*Pc*) across the 9 test trials was calculated separately for each block, and for both the pre-learned and to-be-learned associations. The amount learned in each study block was quantified by taking the first order derivatives of *Pc* values (i.e., the change in performance across study blocks; ΔPc).

Univariate imaging analyses

We specified a first-level general linear model (GLM) to investigate univariate activations associated with encoding processes during learning. Movement parameters derived from the image realignment procedure were included as nuisance regressors and a vector coding the normalized mean white matter intensity per volume was used to account for nuisance fluctuations such as scanner drift and aliased biorhythms. In total, the model included 15 event-related regressors of interest. Twelve of these specified learning events as 6 s boxcar functions grouped according to association type (pre-learned versus to-be-learned) and block (i.e., blocks 1 - 6). The remaining 3 regressors related to test trials which modeled 1) correctly answered to-be-learned tests, 2) incorrectly answered to-be-learned tests, and 3) correctly answered pre-learned test trials as separate event types. Incorrectly answered pre-learned tests were also modeled yet few subjects made pre-learned errors. All test events were specified as delta functions with an onset corresponding to that of the aurally presented cue. An additional regressor of no interest modeled the key presses that followed each test as delta functions. All event-related regressors were convolved with SPM's canonical hemodynamic response function (HRF) and amplitude estimates (β values) were calculated on a voxel-wise basis. Percent signal change was calculated by scaling β values with the corresponding GLM regressor heights and normalizing the resultant values with the GLM constant term.

Prior to statistical analysis, we performed pairwise subtraction contrasts on the to-be-learned and pre-learned β estimates on a block-by-block basis (i.e., to-be-learned > pre-learned). These contrasts were then entered into a second-level mixed-effects model to examine BOLD activations over and above any non-specific block effects not directly related to the to-be-learned trials. The second-level model included random intercepts for each block and participant and two fixed effects predictors; 1) to-be-learned *Pc* values, and 2) to-be-learned *ΔPc* values. The model was estimated with SPM8's nonsphericity modeling algorithms using restricted maximum likelihood estimation (see [37]).

Representational similarity analysis

For the RSA we first estimated multivariate BOLD response to each test trial using a first-level GLM of unsmoothed EPI data. Test events were modeled by unique delta functions and their corresponding *t*-statistics were used to compute the similarity of local BOLD patterns at each point in the brain (described below). Subsequently, RSA contrast matrices were produced to specify predicted changes in representational similarity between pairs of test trials. Two sets of predictions were tested; One relating to the associative learning model (Figure 2B), and another relating to the propose-but-verify model (Figure 2C). Since there were 54 to-be-learned test trials across 6 blocks, each contrast matrix was of size 54 × 54. All diagonal matrix elements were zero-weighted (n = 54). Importantly, elements reflecting correlations between different test blocks were also zero-weighted (n = 1215 in the lower triangle). This ensured that the RSA was not confounded by low frequency noise in the MR signal (e.g., motion, scanner drift). Matrix elements of interest (n = 216 in lower triangle) were mean centered and scaled such that the grand sum was zero and the variance was one.

The RSA was implemented in the CoSMoMVPA toolbox [38]. This involved a searchlight analysis; neural similarity between pairs of test trails was computed at each point in the brain by correlating BOLD patterns within spherical searchlights of a 3-voxel radius (the mean number of voxels per searchlight was 110). This resulted in a 54 × 54 correlation matrix for each brain voxel representing the level of local similarity between all test trails. To compute the agreement between this neural data and the model predictions, each correlation matrix was Fisher-transformed before being multiplied by the RSA contrast matrix under test. We then summed the values within each weighted correlation matrix. This produced a 3D output image representing the total covariance between the neural data and the model predictions at each location in the brain. Note: The Fisher-transformation was used to equate the variances for all possible correlation coefficient to satisfy the assumption of homoscedasticity. Next, the output images were warped to MNI space (as above) before being subject to statistical analysis at the group level. This group analysis involved one-sample t tests that estimated the average size of the RSA effect across participants for each set of model predictions (i.e., the strength of evidence for propose-but-verify versus associative learning). When comparing the goodness-of-fit for each model in a specific brain region, we estimated the strength of association between the neural data and model predictions in a group-wide mixed-effects regression model. The difference in the Bayesian information criterion (BIC) between models was then used to compare their relative goodness-of-fit (see [19]).

Propose-but-verify model

For the propose-but-verify RSA, changes in representational similarity are predicted to follow a stepwise progression; once an association has been verified, BOLD responses to the pseudoword (and recall of the associated object) should decrease in similarity relative to the responses to all other pseudowords. This is because word-object associations are expected to be coded uniquely (i.e., pattern separated) and thus result in distinct activation patterns across voxels. In contrast, prior to any learning, activation patterns in those same voxels should be relatively similar across pseudowords, because unique associations have not been established. We assume that the verification of individual word-object associations occurs in the learning block immediately before the association is first correctly recalled. Note: According to the PbV model by Trueswell et al. [7], verification should only occur on the observation after a hypothesis has been first proposed. As such, it is possible for correct responses to precede verification in some cases. However, given that our experiment included 3 repetitions of each word-object pair per study block, it is most likely that verification preceded correct responses. We represent this dynamic by defining a learning state variable *V* for each association that is 0 prior to the first correct response (pre-verification state), and 1 on and after the first correct response (post-verification state). Given this, the predicted representational dissimilarity between tests of associations *a* and *b* is 0 when both V_a and V_b are 0 (i.e., unverified), and 1 in all other cases:

$$D_{a,b} = \begin{cases} 0, if(V_a + V_b) = 0 \\ 1, otherwise \end{cases}$$

As an example, if the associations for words *a* and *b* were first correctly recalled in the 3rd and 4th test blocks respectively, the hypothesized dissimilarity between *a* and *b* test trials (i.e., $D_{a,b}$) would be zero in test blocks 1 and 2 and one for all others. This dissimilarity measure was then reverse scored and mean centered for all pairwise comparisons of interest before being used in the RSA.

Note: none of the participants in our sample failed to correctly recall any word-object association after verification is assumed to have taken place (that is, once learned, there were no retrieval failures for any to-be-learned word).

Associative learning model

In testing the associative learning model, we make the same assumption as above that learning should cause a decrease in representational similarity between test trials. However, this model suggests that learners encode all co-occurrences and that associations strengthen gradually. As such, the decreases in representational similarity are predicted to proceed gradually and should be related to the ability of each pseudoword to cue a unique object at test. We used a computational model of associative learning in order to estimate changes in associative strength on a trial-by-trial basis [5]. According to the model, learners maintain and adjust associative strengths between all words and objects across learning. Associative strengths are represented by a word x object association matrix, *M*, which is initially empty but filled with a small constant weighting (0.01) when a new word/object is seen for the first time. Associative strengths are updated on each learning event by distributing a constant learning weight, χ, to the subset of co-presented words and objects, *S*. However, χ is not distributed evenly but is preferentially distributed to; 1) word-object pairs with a pre-established association (i.e., prior knowledge), and 2) word-object pairs where the associations of both stimuli are unknown (i.e., uncertain). To quantify this latter uncertainty, a measure of entropy, *H*, is specified to be maximal when a given word (or object) is equally likely to correspond to every other stimulus, and minimal when the association is certain:

$$H_w = -\sum_{o=1}^{n} p(M_{w,o}) \cdot log_2(p(M_{w,o}))$$

Where *w* and *o* are the indices of specific words and objects, respectively, *n* is the number of word-objects pairs, and $p(M_{w,o})$ is the normalized associative strength between word *w* and object *o* across all objects. Note: a similar equation is specified for object uncertainties, H_o. Given this measure of uncertainty, the learning weight (χ) is distributed to elements of *M* based on a scaling parameter λ that governs differential attention to uncertain stimuli versus prior knowledge. Additionally, a trial-by-trial decay parameter, α (constrained between 0 and 1), governs the rate of forgetting:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda(H_w + H_o)} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda(H_w + H_o)} \cdot M_{w,o}}$$

When associations are tested with a word cue, learners are assumed to select an object with a probability that is proportional to the word-object associative strength (see [5]). Because the model has 3 free parameters; α, χ, and λ, we fit these values to each participant by minimizing the sum of the squared error between the model outputs on each test trail across all blocks and the observed behavioral responses. This was done using the "lsnonlin" optimization function in the MATLAB Optimization Toolbox (Mathworks). The means (and *SDs*) of the best fitting parameter values were; $\alpha = 0.85$ (0.040), $\chi = 0.051$ (0.040), and $\lambda = 10.02$ (1.723). The median negative log-likelihood describing the goodness-of-fit for the model was 14.19 (summed across 54 test trails per participant); this corresponds to a modeling accuracy of approximately 76.9%.

Once the model had been fitted, we specified the RSA contrast matrix reflecting predicted changes in representational similarity between pairs of test trials. As test trials presented specific words, the representational similarity between pairs of test events is predicted to be inversely proportional to how specifically they activate unique object representations. Formally, we estimated the representational dissimilarity (*D*) between words *a* and *b* to be equal to the Euclidian distance in their corresponding object associations:

$$D_{a,b} = \sqrt{\sum_{o=1}^{n} (M_{a,o} - M_{b,o})^2}$$

As above, this dissimilarity measure was then reverse scored and mean centered for all pairwise comparisons of interest before being entered into the RSA contrast matrix.

Imaging thresholds

Across all imaging analyses, reported activations survive whole-brain, family-wise error (FWE) corrected thresholds at the cluster-level (cluster defining threshold: $p < 0.001$, one-tailed). Additionally, given our strong *a priori* hypotheses that effects of interest may be observed in the hippocampus, we report activations surviving a small volume correction at the voxel-level within a bilateral hippocampal mask (taken from the AAL Atlas, [39]).

DATA AND SOFTWARE AVAILABILITY

Unthresholded statistical maps for all the reported analyses are available at https://neurovault.org/collections/3002/. Analysis-specific code and data are available on request from the Lead Contact, Sam Berens (sam.berens@york.ac.uk).