Sequence learning creates new cortical representations instead of strengthening initial ones

Kristjan Kalm*, Dennis Norris

MRC Cognition and Brain Sciences Unit, University of Cambridge
15 Chaucer Road, Cambridge, CB2 7EF, UK

E-mail: kristjan.kalm@mrc-cbu.cam.ac.uk.

July 3, 2019

Abstract

We contrast two accounts of how novel sequences are learned. The first is that learning changes the signal-to-noise ratio (SNR) of existing cortical representations by reducing noise or increasing signal gain. Alternatively, learning might cause the initial representations to be recoded into more efficient representations such as chunks. Both mechanisms reduce the amount of information required to store sequences, but make contrasting predictions about cortical activity patterns. We applied representational similarity analysis to patterns of fMRI activity as participants encoded, maintained, and recalled novel and learned sequences of oriented Gabor patches. We fit four models of sequence representation to the activity patterns of novel sequences and tested how the representation changed as a function of learning. We found no evidence for the SNR-change hypothesis. Instead, we observed that in three loci in the occipital and parietal cortex the same sets of voxels encoded both novel and learned sequences but using different encoding schemes. Our results suggest that sequence learning induces recoding rather than simply strengthening the initial representations.

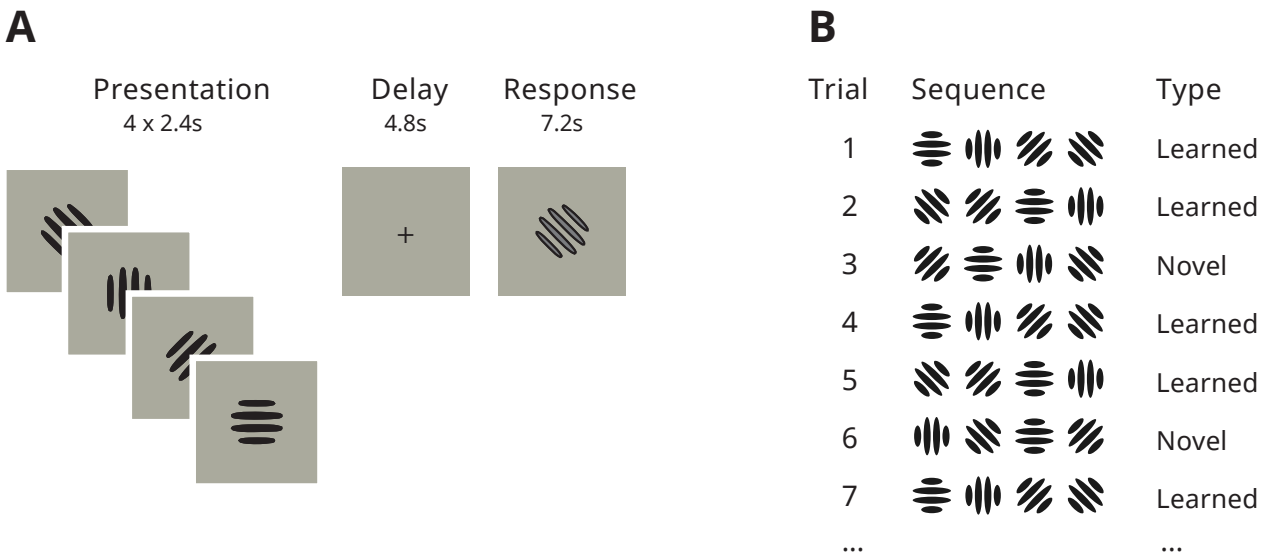
Introduction

The initial encoding of individual events or items into novel sequences is facilitated by the hippocampal formation: humans, primates and rodents with hippocampal damage are unable to retain novel associations between objects and their temporal order (Shimamura, Janowsky, & Squire, 1990; Fortin, Agster, & Eichenbaum, 2002). However, already at this initial encoding stage cortical sequence representations exist in parallel to subcortical ones: neurons encoding the temporal structure of sequences have been observed in monkey prefrontal and parietal regions for novel visual and visuo-spatial sequences (Averbeck, Sohn, & Lee, 2006; Berdyyeva & Olson, 2011; Nieder, 2012), and in motor and premotor areas for novel motor sequences (Crowe, Zarco, Bartolo, & Merchant, 2014; Carpenter, Georgopoulos, & Pellizzer, 1999; Merchant, Pérez, Zarco, & Gámez, 2013). Furthermore, humans and animals with hippocampal lesions are unable to learn new sequences but are able to recall already learned ones (Mayes et al., 2001; Devito & Eichenbaum, 2011). This suggests a dissociation between the hippocampal and cortical representations: both are created simultaneously whilst affected differently by learning. Consolidation models of human learning propose that the initial hippocampal-dependent associative memories are recoded as they undergo consolidation resulting in more efficient cortical representations (McClelland, McNaughton, & O'Reilly, 1995; Treves & Rolls, 1994; McKenzie et al., 2016). However, it is unclear how this recoding happens for novel sequences, what is the recoding computation, and how are the neural codes for novel and learned sequences exactly different. Although studies have informed us that learned sequences are eas-

ier to decode from cortical activation patterns compared to novel ones (Pinsard et al., 2019; Wiestler & Diedrichsen, 2013) the nature of the learning-induced change remains unclear.

In this paper we study how learning changes the cortical representations of sequences. We use fMRI data from a human sequence memory task (Fig 1) to to identify which cortical areas encode novel visual sequences by fitting computational models of sequence representation to fMRI activation patterns. In regions which encode novel visual sequences we contrast the cortical activation patterns of novel with those of learned sequences to identify the change in neural representation that accompanies learning. Specifically, we tested for two contrasting hypotheses.

Figure 1: Task. (A) Single trial: participants had to recall a sequence of four Gabor patches in the order it was presented after a delay. The size of the stimuli within the display area is exaggerated for illustrative purposes. (B) Trials types and progression: all recalled sequences were either learned beforehand or novel. See *Task* in *Methods* for details.

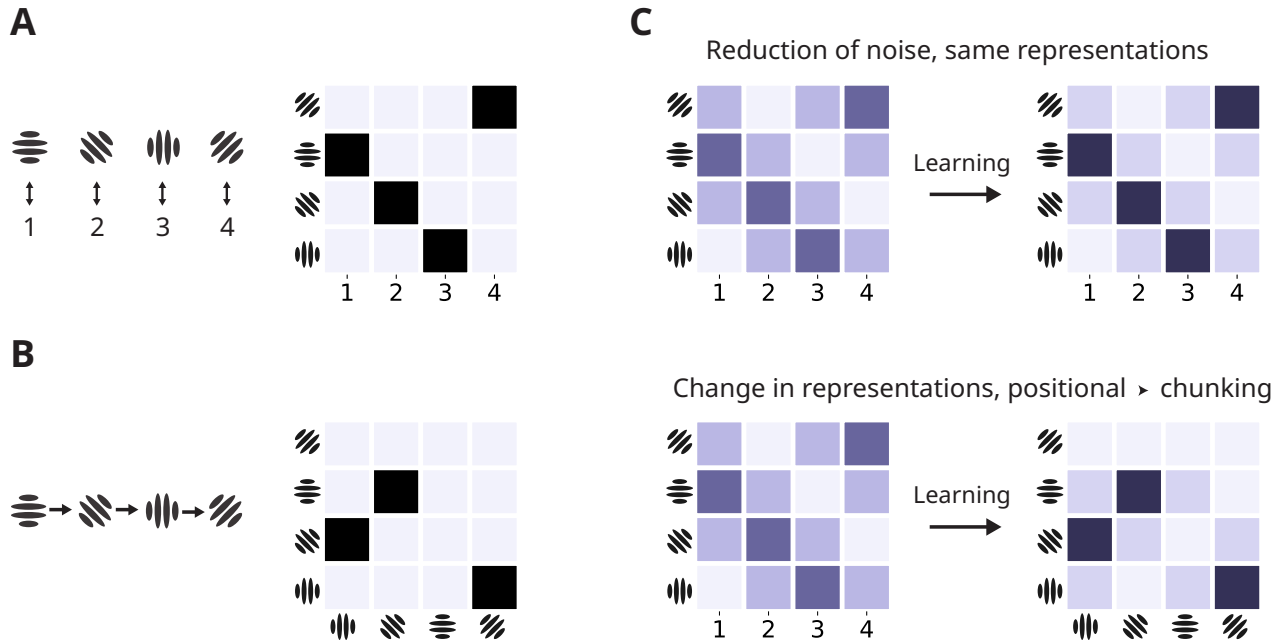


The first hypothesis posits that learning strengthens the existing cortical representations by changing the signal-to-noise ratio (SNR) of neural responses over repeated exposures (Dosher & Lu, 1998; Eldar, Cohen, & Niv, 2013). For example, several days of motor sequence learning have been shown to make the cortical representations of learned sequences more readily distinguishable from each other compared to untrained sequences (Wiestler & Diedrichsen, 2013).

Alternatively, repeated presentations might elicit the recoding of cortical representations using more efficient codes that utilise the statistical regularities in the stimuli. For example, learned sequences may be recoded into chunks of successive items (Gobet et al., 2001; Du & Clark, 2017). A chunked sequence is computationally more efficient as it allows binding individual items to each other to form a single unit (Fig 2).

The crucial difference between these two learning approaches is that in the first case the sequence codes remain the same, whilst in the latter case a new and more efficient code emerges.

Figure 2: Change in sequence representation by learning. Four oriented Gabor patches in a sequence: (A) Positional representation: items (Gabors) are associated with an ordinal position in a sequence. (B) Chunk representation: items are associated with each other forming chunks of item pairs. Here chunks are pairwise item-item associations (2-grams). (C) Change in representations by learning. Top: learning reduces noise in representations. Matrices display the association strengths. Bottom: learning, changes representations from position-item associations to item-item chunks.



Importantly, both learning mechanisms reduce the amount information required to represent a sequence (Fiser, 2009; Fiser, Berkes, Orbán, & Lengyel, 2010) and therefore are hard to dissociate on the basis of behavioural measures. In this study we take advantage of the fact that the two learning mechanisms necessarily predict contrasting neural response pattern similarities between novel and learned sequences.

We show that, in line with previous studies, activation patterns in the parietal and visual cortices represent novel visual sequences (Pinsard et al., 2019; Kornysheva & Diedrichsen, 2014; Wiestler & Diedrichsen, 2013). We found no evidence for the hypothesis that learning changes noise levels in those sequence representations. Instead, we observed that in three loci in the occipital and parietal cortex the same sets of voxels encoded both novel and learned sequences but using different encoding schemes.

Results

Behavioural measures

In our task participants manually recalled a four-item sequence of Gabor patches after a 4.8s delay period (Fig 1). The sequences were either novel or learned. We calculated two behavioural measures of recall for both types of sequences: how many items were recalled in the correct position, and the average time between consecutive key presses. The proportion of correctly recalled items was roughly the same for novel and learned sequences: 0.96 vs. 0.97, with no significant difference across subjects ($p = 0.22, df = 21$). This was expected since both novel and learned sequences were four items long and should therefore fit within participants short term memory span. However, participants were consistently faster in recalling learned sequences: the average time between consecutive key presses was 0.018 seconds shorter for learned sequences ($t = -3.04, p = 0.007, df = 21$). This shows that novel and learned sequences were processed differently by participants.

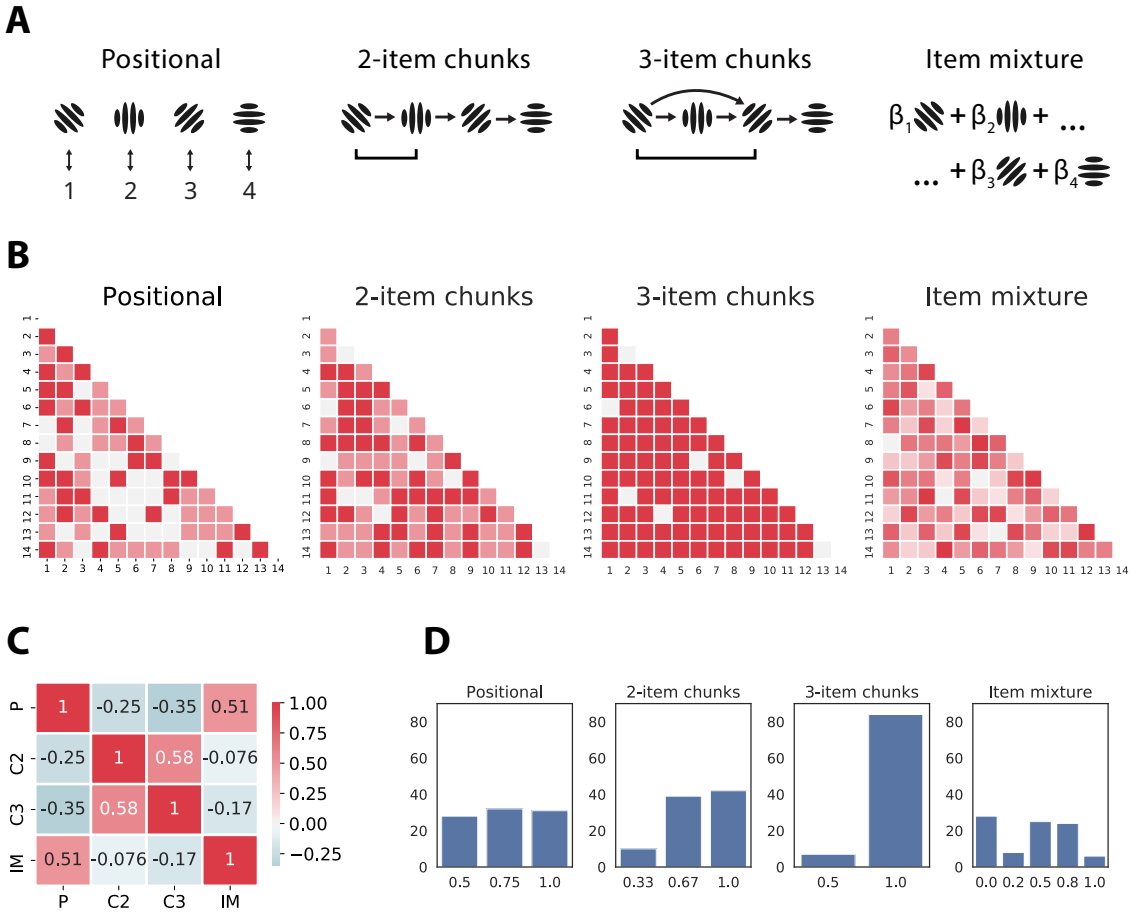
Cortical representation of novel sequences

Based on two broad classes of computational models we specified four alternative models of sequence representation (see *Models of sequence representation* in *Methods*). The four models either encode novel sequences as associations between items and temporal order positions (*positional model*, Fig 3A-B, first column), pairs of successive item-item associations (*2-item chunks*, Fig 3A-B, second column), triplets of successive between-item associations (*3-item chunks*, Fig 3A-B, third column), or as a weighted mixture of item representations (*item mixture model*, Fig 3A-B, fourth column). The item mixture model is a control model which tests for a null-hypothesis that instead of sequence representations voxel activity could be better explained by the superposition of patterns for constituent individual items in the sequence (see e.g. Yokoi, Arbuckle, & Diedrichsen, 2018).

We parcellated the dorsal visual processing stream bilaterally into 74 anatomically distinct regions of interest (ROI). In every ROI we fit the four sequence representation models to voxel activation patterns corresponding to novel sequences. The model fits showed how well each model predicted the distance between pairs of voxel activity patterns corresponding to individual novel sequences (see *Representational similarity analysis* in *Methods*). This was done separately for all three task phases (presentation, delay, response).

Besides identifying the best-fitting model we sought to determine whether any of the models also accounted for the representational structure observed across participants. For this purpose we determined the noise ceiling for the model fits in each of the ROIs. The noise ceiling is the expected fit with data achieved by the theoretical true model given the noise (see *Noise*

Figure 3: Four models of sequence representation. (A) Visualisation of how different models represent a four-item sequence. For chunking models the bracket underneath the items illustrates a single chunk in a sequence. (B) Similarity of 14 individual sequences presented in the task of this study as estimated by respective sequence representation models. (C) Between-model similarity measured as Pearson's correlation coefficient between model-produced pairwise distance matrices for the fourteen individual sequences (panel B). P: Positional, C2: 2-item chunks, C3: 3-item chunks, IM: Item mixture. (D) Distribution of normalised pairwise similarity distances per model. The spread of the distance values indicates the predictive power of the model.



ceiling estimation in *Methods*). Any model which does not reach the noise ceiling should not be considered as a plausible explanation of voxel responses.

Only the positional model and the item model reached the noise ceiling for novel sequences

For novel sequences we found that only the positional model and the item model reached the noise ceiling for novel sequence representation in 18 regions out of 74 ROIs bilaterally (Table 1). These were the only models to reach the noise ceiling – no chunking model reached the lower bounds of the noise ceiling in any ROIs. This suggests that the chunking models are not a good fit for novel sequence representations anywhere in the dorsal visual processing stream. Since

the predictions of the positional and chunking models are inversely correlated, such relationship between the model fits was expected: both cannot be positively correlated with the same data (see Fig 3C and *Models of sequence representation* in *Methods*).

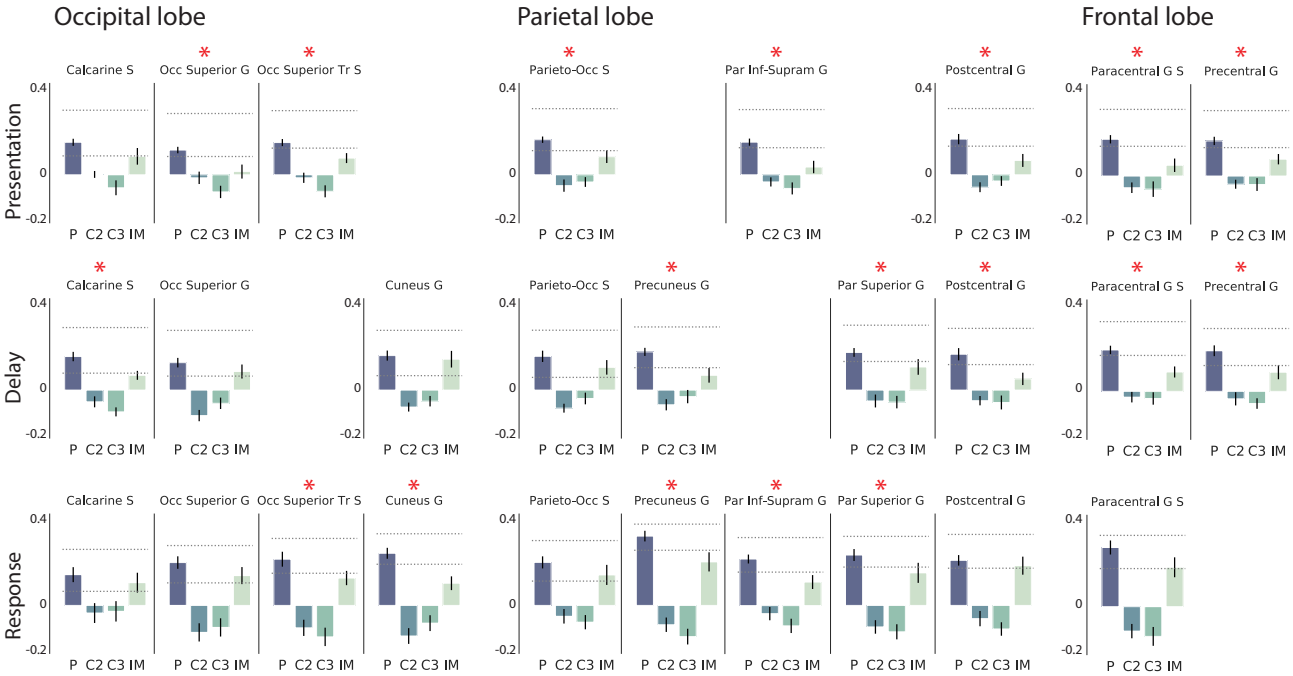
Table 1: Evidence for the representation of novel sequences. Anatomical region suffixes indicate gyrus (G) or sulcus (S). Asterisks (∗) represent the positional model and daggers (†) the item model reaching the lower bound of the noise ceiling. Regions in which the item model reached the noise ceiling are displayed in italics.

Lobe	Name	Presentation	Delay	Response
Frontal	Central S		∗	∗
Frontal	*Frontal Middle G*	∗	†,∗	
Frontal	*Paracentral G S*	∗	∗	†,∗
Frontal	Precentral G	∗	∗	
Occipital	*Calcarine S*	∗	∗	†
Occipital	*Cuneus G*		†,∗	∗
Occipital	*Occipital Middle G*	∗	†,∗	
Occipital	*Occipital Superior G*	∗	†	†,∗
Occipital	Occipital Superior Transversal S	∗		∗
Occipital	*Parieto-Occipital S*	∗	†,∗	†,∗
Parietal	Intraparietal Posterio-Transversal S			∗
Parietal	*Parietal Inferior-Angular G*		†,∗	∗
Parietal	Parietal Inferior-Supramarginal G	∗		∗
Parietal	Parietal Superior G		∗	∗
Parietal	*Postcentral G*	∗	∗	†,∗
Parietal	Postcentral S	∗		∗
Parietal	Precuneus G		∗	∗
Temporal	Temporal Superior S			∗

The item model reached the noise ceiling in several visual, motor, and parietal regions for all task phases (Fig 4, Table 1). The item model reflects a control hypothesis that instead of sequence representations voxel patterns might instead reflect a mixture of item representations. In line with a previous study (Yokoi et al., 2018) we observed that the item model could explain the voxel activity for the response phase of the task in the motor regions (postcentral and paracentral gyri, Fig 4) suggesting that this activity represented the superposition of individual finger movements during the manual response. The finger-movement explanation is further supported by the finding that the item model did not reach the noise ceiling in the delay or presentation phases of the task in those motor regions.

The item model also reached the noise ceiling in several visual and parietal regions. Notably, this only happened in the the delay and response phases and not in the presentation phase. The evidence for the item model across the whole dorsal processing stream in the delay and response phases supports recent findings that both visual and motor representations of individual items

Figure 4: Model fits in the dorsal visual processing stream. Overview of the ROIs where any of the model fits reached the noise ceiling. Red asterisks mark regions where the fit with the positional model was significantly greater compared to the item model ($p < 10^{-3}$) and the item model did not reach the noise ceiling. Rows represent task phases (presentation, delay, response), columns individual ROIs in the dorsal visual processing stream. Y-axis displays the average Spearman correlation coefficient, error bars represent the standard error of the mean (SEM). Dashed lines in bar plots represent the lower and upper bounds of the noise ceiling. Bar legends: P: Positional, C2: 2-item chunks, C3: 3-item chunks, IM: Item mixture.



are concurrently engaged in working memory guided tasks in both delay and response phases (van Ede, Chekroud, Stokes, & Nobre, 2019).

Since we cannot rule out the possibility that the regions where the item model reached the noise ceiling are not engaged in item rather than sequence representation we excluded those ROIs from further analysis (region names displayed in italics in Table 1). It should also be noted that the positional model and item model are *a priori* correlated to each other ($\rho = 0.52$, Fig 3C) since item position effects dominate item mixture components (see *Item mixture model* in *Methods*). This means that in regions with significant correlation with the positional model we should also always see some positive correlation with the item model. However, here we required that the item model should not reach the noise ceiling and be significantly less correlated to data than the positional model in order to rule out the null hypothesis that the region is representing individual items or finger movements.

The positional model is the only plausible model for novel sequences

In nine regions in the dorsal visual processing stream the positional model was not only the best-fitting and sufficient (reaching the noise ceiling) model for novel sequences but also the only model which reached the noise ceiling (Fig 4, Table 1). This finding is consistent with previous fMRI studies which have shown that the same regions in the occipital and parietal cortices are reliably sensitive to the memoranda and response actions in visual working memory tasks (Bettencourt & Xu, 2015; Ester, Sprague, & Serences, 2015; Christophel, Hebart, & Haynes, 2012).

However, there were significant differences in positional model fits for different phases of the task. For the presentation phase (9.6s duration, four differently oriented Gabors, Fig 1) the positional model predicted voxel patterns not only in the visual and parietal cortices but also in primary and association motor areas (Fig 4, top row). However, none of the motor regions encoded novel sequences during the response phase despite the significant evidence of sequence representation in the presentation and delay phases (Fig 4, three rightmost columns). In those ROIs the fit with the positional model was not significantly different from the item model during the response phase (Fig 4, bottom row) suggesting that individual finger movements dominated the voxel patterns in the motor areas.

In sum, our analysis of novel sequence representation shows that combining fMRI data with representational similarity models allows more precise quantification of the relationship between cortical activity patterns and models of sequence representation. Using an explicit control model we were able to delineate neural sequence representations from correlated representations of individual items. For novel visual sequences only the positional model provides a plausible explanation of data in the dorsal visual processing stream. Our findings are in line with previous research that novel sequences are initially encoded in terms of associations between items and their temporal positions in both animal (Fig 2A; Berdyyeva & Olson, 2010, 2011; Averbeck et al., 2006; Ninokura, Mushiake, & Tanji, 2004) and human cortex (Heusser, Poeppel, Ezzyat, & Davachi, 2016; Hsieh, Gruber, Jenkins, & Ranganath, 2014; Kalm & Norris, 2014). Several regions in the dorsal stream were also encoding individual item or finger movement representations and those regions were excluded from the analysis of learning effects.

Comparison of learning mechanisms

The analysis of learning effects was carried out in the 9 ROIs where we found significant and sufficient evidence for novel sequence representation across task phases (Table 1), as the effects of learning on sequence representation can only be established in regions where such representations have been observed. We tested the data from these regions for two alternative hypothesis: (H1) learning reduces noise in sequence representations, or (H2) learning recodes

sequences.

H1: Learning reduces noise in sequence representations

If learned sequences are represented similarly to novel sequences but with less noise, then the activity patterns elicited by learned sequences should be similar to those elicited by novel sequences, and this similarity should, on the average, be greater than within novel sequences (see *Representational similarity analysis* in *Methods* for the detailed derivation of the hypotheses).

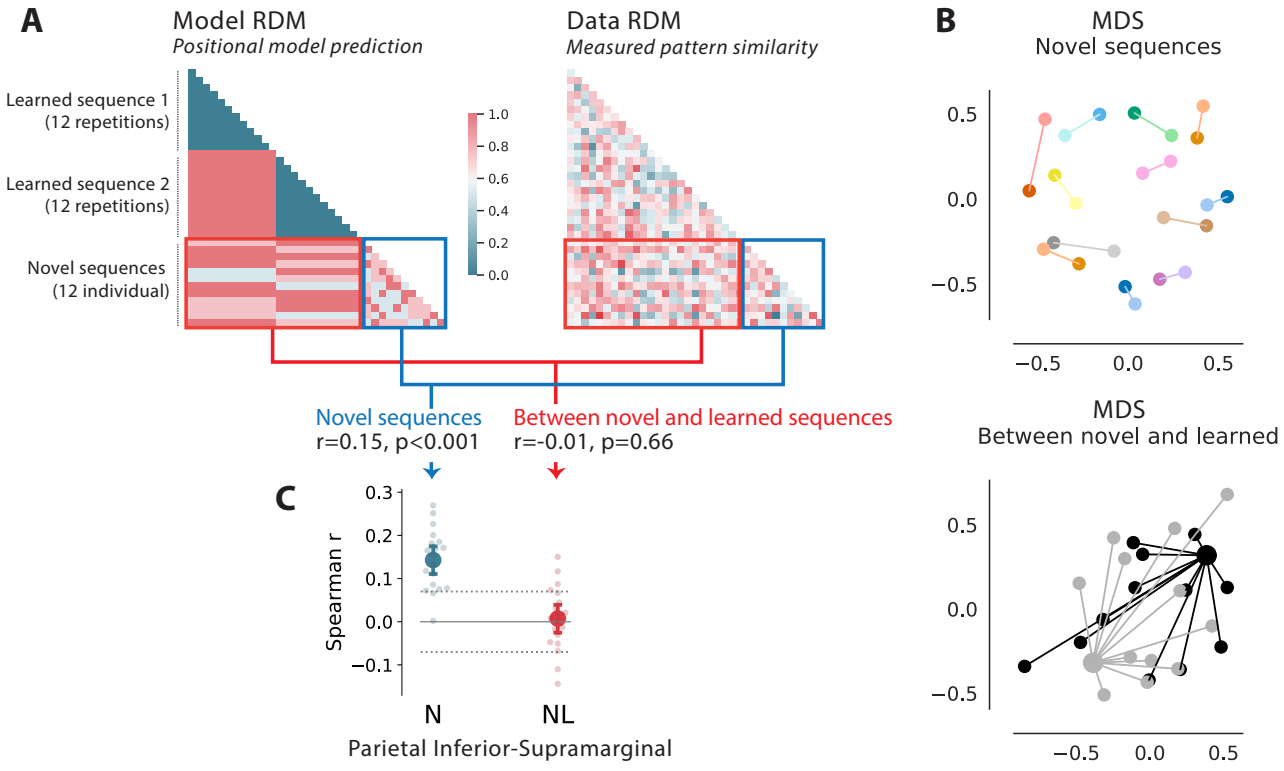
We found no evidence for the noise reduction hypothesis in any of the ROIs that encoded novel sequences. Specifically, there was no significant correlation between the predictions made by the positional model and the voxel pattern similarity between learned and novel sequences (Fig 5A illustrates graphically how this was measured). In all of the 9 regions where the positional model predicted similarity between novel sequences it failed to predict the similarity between novel and learned sequences (Fig 5C).

We can further illustrate this result by visualising the fit between model predictions and data by projecting the representational similarity onto 2D by using multidimensional scaling (MDS). Fig 5B shows how well the positional model predicts similarity between novel sequences (top panel) and between novel and learned sequences (bottom panel). If learned sequences were also encoded positionally but with less noise, then the distances between model predictions and data should be shorter for the novel-learned comparison (bottom) than for novel sequences only. Fig 5B shows how the opposite is true for the voxel patterns in the parietal inferior-supramarginal gyrus (same scale on both plots).

H2: Learning recodes sequences

If learned sequences are represented in areas which encode novel sequences (Table 1), but not positionally, then we should be able to detect these learned sequence representations by using a decoding approach. By using voxel patterns corresponding only to the two learned sequences (*learned sequence A* vs. *learned sequence B*) we used a pattern classification approach to test whether those voxels encode the learned sequences (see *Linear classification of two learned sequences* in *Methods*). However, to ensure that the representation of novel and learned sequences were not dependent on two separate subsets of voxels in the same ROI, we constrained our decoding analysis so that for a single subject only those voxels which had a significant fit with the positional model were included in the classification analysis. This was achieved by running the novel sequence representation RSA (exactly as described above) using a searchlight approach resulting in individual correlation scores for every voxel in the ROI. In other words, in every ROI we only ran the decoding analysis on a subset of voxels which showed significant evidence for the positional model.

Figure 5: Novel vs learned sequence representations. (A) Predictions of the positional model for the novel sequences (blue squares connected with blue lines) and for the similarity between novel and learned sequences (red rectangles connected with red lines). The blue squares contain the similarities between all novel sequences. The red rectangles contain the similarities between novel sequences and the two learned sequences. Here we use the latter to test for alternative learning hypotheses (panel C). (B) Multidimensional scaling (MDS) of the representational similarity depicting in 2D the average distances between the positional model predictions and data from the parietal inferior-supramarginal gyrus. Top figure shows positional model predictions for individual novel sequences with lines joining predictions to observed data (the shorter the line the closer data is to the prediction). Bottom figure shows the positional model's predictions for the distances between novel and two learned sequences. The predictions for the two learned sequences are depicted with large dots (black and grey) whilst the data are marked with smaller dots. (C) Model predictions and fit with data in the parietal inferior-supramarginal gyrus: while the positional model significantly predicts the similarity between novel sequences (blue) it performs at chance level when predicting the similarity between novel and learned sequences (red). Dotted lines show significance thresholds for positive and negative correlation with the model predictions ($df = 21$). Dots: individual subjects' values; error bars around the mean represent bootstrapped 95% confidence intervals.



Of the regions which encoded novel sequences (Table 1) we found significant evidence for learned sequence representation in three brain regions: the parietal inferior-supramarginal gyrus, the postcentral sulcus, and the occipital superior transversal sulcus (Table 2, Fig 6). The classification results were only significant for the presentation phase of the task; we could not decode between two learned sequences during either the delay or the response phases. This suggests that the decoding results do not reflect individual finger movements. Similarly, if the decoding analysis reflected a combination of voxel patterns for constituent individual items

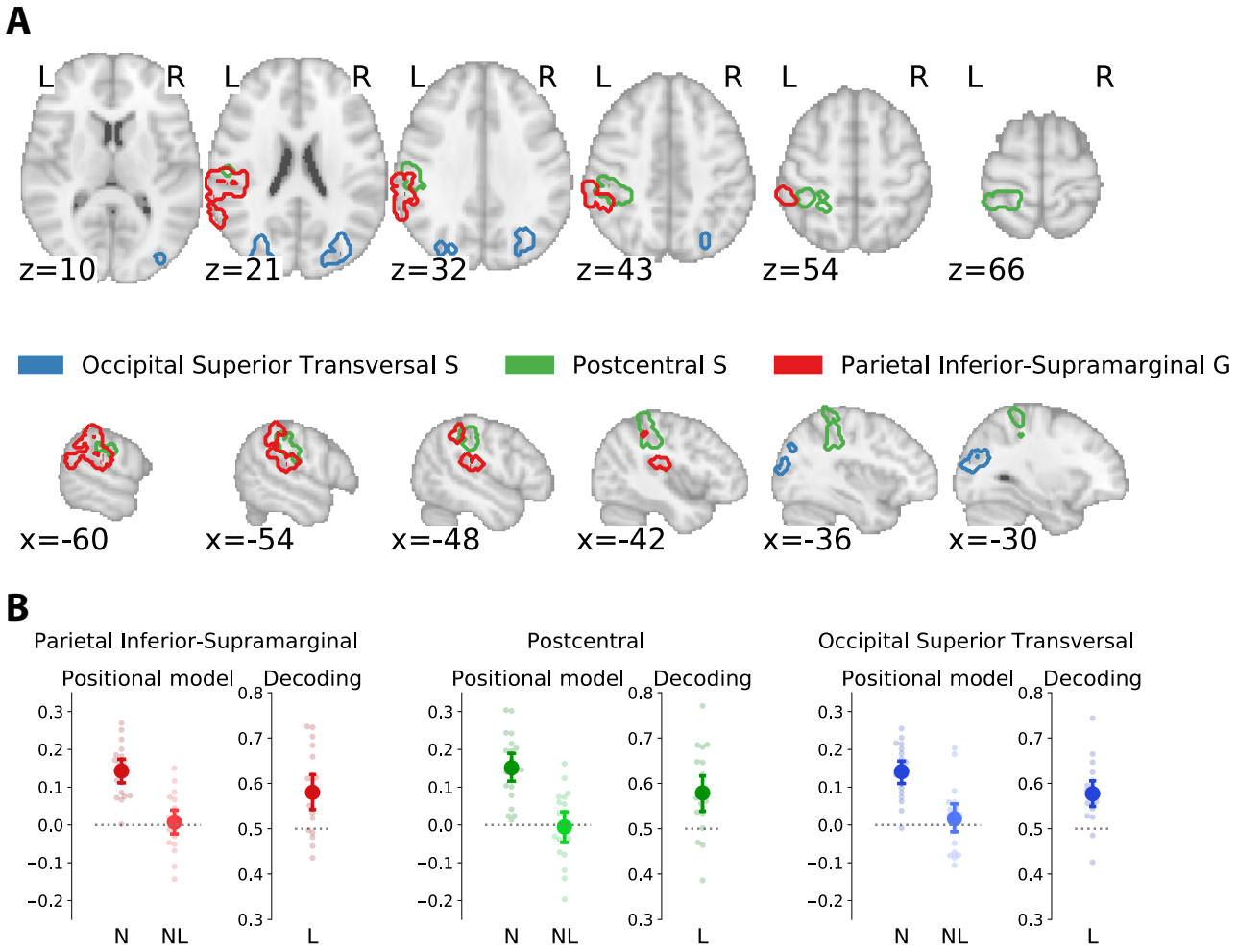
instead of sequence representations, then we should have seen a significant fit with the item model and novel sequences for these regions (Table 1, Fig 4). Instead, the item model did not reach the noise ceiling in any of these regions.

Table 2: Decoding learned sequences. Novel-sequence encoding brain regions where the classification accuracy for two learned sequences was significantly above chance across participants. Anatomical region suffixes indicate gyrus (G) or sulcus (S).

Lobe	Region	t-value	p
Parietal	Parietal Inferior-Supramarginal G	4.085	0.00069
Parietal	Postcentral S	4.015	0.00074
Occipital	Occipital Superior Transversal S	3.904	0.00104

Our results show that those three cortical regions – the parietal inferior-supramarginal gyrus, the postcentral sulcus, and the occipital superior transversal sulcus – encode both novel and learned sequences (Fig 6B). The representation of novel sequences in all three is positional, reflecting association of items to their temporal order positions in a sequence. However, the same set of voxels representing the novel sequences also encode learned sequences, but not positionally, indicating a change in the representational code. These results suggest that instead of strengthening the initial representations sequence learning proceeds by recoding the initial stimuli using a different, and potentially more efficient set of codes.

Figure 6: Anatomical localisation of learning effects. (A) Regions which encode both novel and learned sequences. Regions are shown for a single participant (P-9) in the MNI152 standard space. Red: the parietal inferior-supramarginal gyrus; green: the postcentral sulcus; blue: the occipital superior transversal sulcus. Top: axial slices; bottom: saggital slices, left hemisphere. (B) Evidence for the representation of novel and learned sequences: plotted for every ROI are the average RSA correlation scores for novel sequences (N) and between novel and learned sequences (NL), and the average decoding accuracy between two individual learned sequences (L). Dots: individual subjects' values; error bars around the mean represent bootstrapped 95% confidence intervals.



Discussion

In the current study we contrasted two hypotheses about how cortical representations of sequences change as a result of learning. First, we considered the possibility that repeated presentations might change the signal-to-noise ratio of existing representations by means of either a reduction in noise or an increase in signal gain. Alternatively, repeated presentations might lead the initial representations to be recoded into more efficient representations such as chunks. At the behavioural level both mechanisms would have the effect of improving performance in a recall task. However, the two accounts make different predictions about changes in the pattern of neural representations with learning. We were able to test these contrasting predictions by comparing the patterns of activity elicited by novel and learned sequences with those predicted by sequence representation models.

Learning induces recoding of cortical representations

We found that *novel* visual sequences were represented as position-item associations in a number of anatomically distinct regions in the dorsal visual processing stream. This is in line with previous research reporting that initial hippocampal-dependent sequence representations are associative, binding individual events to a temporal order signal which specifies their position in a sequence (Hsieh et al., 2014). However, we found no evidence that these brain regions also represented *learned* sequences positionally. Instead, we observed that learning proceeds by recoding the initial stimuli using a different, and potentially more efficient set of codes.

This raises an obvious question – if not positional, then what is the encoding model of learned sequences? Specifically, could they be encoded as chunks, an encoding scheme frequently reported in the behavioural literature (Gobet et al., 2001; Orban, Fiser, Aslin, & Lengyel, 2008)? Unfortunately, the experimental design did not allow us to explore this question: here we used only two learned sequences, which made it impossible to perform RSA over learned sequences similar to the novel ones (we used 12 individual novel sequences). Restricting the number of learned sequences was necessary for several reasons: the difficulty of learning multiple overlapping sequences, the necessary dissimilarity between learned and novel sequences, and the limited number of possible four-item sequences (see *Sequence generation* in *Methods* for details). In short, the study was optimised for detecting the representational scheme of novel sequences and then testing for a learning-induced change in representation. Further research is required to elucidate the nature of cortical representations of learned sequences, however several previous studies have suggested that in the subcortical regions learned sequences are encoded using chunk-like codes: learned motor sequences as chunks in the dorsolateral striatum (Graybiel, Aosaki, Flaherty, & Kimura, 1994), and odours as chunks in the ventral basal ganglia (Greve & Fischl, 2009).

Several recent studies of motor skill learning have reported better classification-based measures for learned finger sequences compared to novel ones in similar cortical regions: for trained vs. untrained sequences in motor and parietal cortices (Wiestler & Diedrichsen, 2013), and in prefrontal and premotor areas (Pinsard et al., 2019). However, these studies have just compared the discriminability of individual learned sequences to novel ones and therefore cannot inform us what is the nature of the change facilitated by learning. Here we show for the first time that learning visual sequences induces recoding of initial representations: a common hallmark of human learning observed in a range of tasks from different modalities at multiple time scales (see Fiser et al., 2010 for a review).

Our findings also support a model of the hippocampal-cortical learning system where the initial sequence representations are associative, episodic, and created by single exposure (Buzsáki & Tingley, 2018; Legéndy, 2017; Kesner & Rolls, 2015). We show that parallel to the hippocampal codes the initial cortical representations of visual sequences are also associative, binding individual items to their positions in a sequence. This suggests that they are created in parallel, or that the hippocampal representations act as associative pointers to the cortical representations (Legéndy, 2017). Further research focussing on the direct comparison between the hippocampal and cortical sequence representations is necessary to shed light on this relationship. In this study we were unable to compare the hippocampal sequence representations to the cortical ones since our anatomical 'coverage' of the brain provided by the MRI scanner was optimised for the visuo-motor cortical areas and therefore did not reach the deep subcortical regions (efficient subcortical MRI data acquisition usually requires dedicated tuning).

Initial sequence representations might be suboptimal for long-term storage

Recoding initial sequence representations is also advantageous from an optimal learner perspective. A single sequence can readily be represented as a set of position-item associations. Each item can be uniquely associated with a single position, and this will be sufficient to recall the items in their correct order. However, such a coding scheme fails when the requirement is to learn multiple overlapping sequences. The sequences $ABCD$ and $BADC$ cannot be learned simultaneously simply by storing position-item associations, as the resulting set of associations would be equally consistent with the unlearned sequence $ABDC$ – a different coding scheme is therefore required for long-term learning. In contrast, simple paired-associate learning (pig-book, rain-leg) might be able to rely simply on strengthening associations and not require the development of new codes. However, most naturally occurring sequences (words or sentences in a language, events in complex cognitive tasks live driving a car or preparing a dish) are not made out of items or events which only uniquely occur in that sequence. Hence a sequence learning mechanism has to be able to learn multiple sequences which are re-orderings of the same items. Strengthening of position-item associations would result in significant interference

between the learned associations. Our findings are also consistent with behavioural data on memory for sequences where there is evidence for the use of positional coding when recalling novel sequences (Fischer-Baum & McCloskey, 2015) while learned sequences show little indication of positional coding (Cumming, Page, & Norris, 2003). Furthermore, sequence learning is still possible when the position of items (but not their order) changes from trial to trial (Kahana, Mollison, & Addis, 2010).

Conclusions

Our results show that sequence learning proceeds by recoding cortical representations. Although the initial hippocampal-dependent associative representations of novel sequences may be sufficient to support immediate recall, multiple sequences can only be learned by developing higher order representations such as chunks. Our findings show that such recoded representations of learned visual sequences can be found in the occipito-parietal cortical regions.

Methods

Models of sequence representation

Decades of psychology research have yielded a powerful set of computational models for explaining how sequences are represented in the brain (see Hurlstone, Hitch, & Baddeley, 2014 for a review). Such models can be assigned to two broad classes dependent on whether they (1) associate items to each other, or (2) associate items to some context signal specifying their position in a sequence (Fig 2).

In the first case sequences are encoded by associating successive items to each other and the association weights are usually expressed in terms of transitional probabilities (Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015). Examples of this class of models are chaining models (Murdock, 1997), chunking models (Gobet et al., 2001; Gobet, Lloyd-Kelly, & Lane, 2016), and n-gram models (Teh, 2006). Independent of how the model is implemented (e.g. neural network, process model) the underlying representation of sequences relies on item-to-item associations. In the second case sequences are formed by associating items to some external signal specifying the temporal context for sequences. This context signal can be a gradually changing temporal signal (Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1992; Hasselmo, Howard, Fotedar, & Datey, 2006), a discrete value specifying the item's position in a sequence (Lee & Estes, 1981), or a combination of multiple context signals (Henson, 1999, 1998). Again, common to all of these models is the underlying association of item representations to the temporal order signal. These common representational features within model classes can be utilised to generate predictions of between-sequence similarity.

Positional model

We first define a model where items are associated with a temporal order context signal. We call this conventionally a 'positional model' (see Henson & Burgess, 1997) as items are associated with their 'positions' in a sequence.

When sequences are represented as position-item associations they can be described in terms of their similarity to each other: how similar one sequence is to another reflects whether common items appear at the same positions. Formally, this is measured by the Hamming distance between two sequences:

$$D_H(S_i, S_j) = \sum_{i=1}^{k} |x_i - y_i| \tag{1}$$

so that

$$x_i = y_i \Rightarrow 0$$
$$x_i \neq y_i \Rightarrow 1$$

where x_i and y_i are the i-th items from sequences S_i and S_j of equal length k respectively.

Consider two sequences $\{A, B, C, D\}$ and $\{C, B, A, D\}$: they both share two item-position associations (B at the second and D at the fourth position) hence the Hamming distance between them is 2 (out of possible 4). The Hamming distance can therefore be used as a model of sequence representation in the brain: if sequences are coded as item-position associations then the similarity of neural activity patterns elicited by sequences should follow the Hamming distance. Fig 3B (left panel) shows the similarity between all individual sequences as measured by the Hamming distance.

We use the between-sequence similarity as defined by the Hamming distance as a positional model's prediction about the similarity between fMRI activation patterns. This allows us to test whether a particular set of voxels encodes information about sequences using a positional encoding scheme. *Representational similarity analysis* in *Methods* provides the details of the implementation.

Chunking model

Again, following convention, here we call the second class of models, where items are associated to each other, 'chunking models'. A chunk is a set of consecutive items in a sequence and can hence be defined via item-item associations as opposed to item-position associations (Fig 2B). Here we represent chunks as *n-grams* – for example, a four-item sequence $\{A, B, C, D\}$ can be

unambiguously represented by three 2-grams AB, BC, CD so that every 2-gram represents transitional probabilities between successive items and the whole sequence as a first order Markov chain:

$$p(A)p(B|A)p(C|B)p(D|C).$$

Note that we can similarly use 3-grams, 4-grams or suitable combinations as any n-gram can be expressed as $n-1$ order Markov chain of transitional probabilities.

The probabilistic representation of chunks can be used to derive a hypothesis about the similarity between chunked sequences: the between-sequence similarity is proportional to how many common chunks they share. For example, the sequences FBI and BIN are similar from a 2-gram chunking perspective since both could be encoded using a 2-gram where B is followed by I (but share no items at common positions and are hence dissimilar in terms of item-position associations). This allows us to define a pairwise sequence similarity measure which counts how many n-grams are retained between two sequences:

$$D_C(S_i, S_j) = |C_i \cap C_j|$$

where C_i and C_j are the sets of n-grams required to encode sequences S_i and S_j respectively. All possible constituent n-grams of both sequences can be extracted iteratively starting from the beginning of sequence and choosing n consecutive items as an n-gram:

$$D_C = \sum_{i=1}^{k-(n-1)} \left| \begin{pmatrix} x_i \\ \vdots \\ x_{i+(n-1)} \end{pmatrix} \neq \begin{pmatrix} y_i \\ \vdots \\ y_{i+(n-1)} \end{pmatrix} \right|$$

where x_i and y_i are the i-th items from sequences S_i and S_j of length k respectively. Effectively, the chunking similarity D_C counts the common members between two chunk sets. Given sequence length k this similarity can accommodate chunks of all sizes n (as long as $n \leq k$). To make the measure comparable for different values of n we need to make the value proportional to the total number of possible n-grams in the sequence and convert it into a distance measure by subtracting it from 1:

$$D_C = 1 - \gamma \sum_{i=1}^{k-(n-1)} \left| \begin{pmatrix} x_i \\ \vdots \\ x_{i+(n-1)} \end{pmatrix} \neq \begin{pmatrix} y_i \\ \vdots \\ y_{i+(n-1)} \end{pmatrix} \right| \tag{2}$$

where γ is a normalising constant:

$$\gamma = \frac{1}{k-(n-1)}.$$

Effectively, the chunking distance D_C counts the common members between two chunk sets. We then used the 2-gram and 3-gram distance measures to derive sequence representation predictions for 2-item and 3-item chunking models (see *Representational similarity analysis* below).

The prediction made by the chunking distance D_C is fundamentally different from the prediction made by the Hamming distance D_H (Eq 1): the chunking distance assumes that sequences are encoded as item-item associations (Fig 2B) whilst the Hamming distance assumes sequences are encoded as item-position associations (Fig 2A). Fig 3B shows the similarity between individual sequences according to two chunking models: sequences represented as 2-grams and 3-grams.

Item mixture model

This model is not a sequence representation model but rather an alternative hypothesis of what is being captured by the fMRI data. This model posits that instead of sequence representations fMRI patterns reflect item representations overlaid on top of each other like a palimpsest, so that the most recent item is most prominent. For example, a sequence $\{A, B, C, D\}$ could be represented as a mixture: 70% the last item (D), 20% the item before (C), and so forth. In this case the mixing coefficient increases along the sequence. Alternatively, the items in the beginning might contribute more and we would like to use a decreasing mixing coefficient. If all items were weighted equally the overall representations would be identical as each sequence is made up of the same four items. Formally, we model an item mixture representation M of a sequence as a weighted sum of the individual item representations:

$$M = \mathbf{I}\beta \tag{3}$$

where $\mathbf{I}$ is the four-dimensional representation of individual items in the sequence and β is the vector of mixing coefficients so that β_n is the mixing coefficient of the n-th item in $\mathbf{I}$ so that

$$0 < \beta_n \leq 1, \text{ and } \sum_{m=1}^{N} \beta_n = 1.$$

where N is the length of the sequence. The rate of change of β (to give a particular β_n a value) was calculated as

$$\beta_n = \alpha\beta_0(1-\theta)^n,$$

where θ is the rate of change and α normalising constant. In this study we chose the value of θ so that $\beta = \{0, 0.1, 0.3, 0.6\}$. Although different values of θ create slightly different mixture models they all have the desired effect of representing a combination of individual items rather

than item-item or item-position associations.

Distances between two item mixture representations M_i and M_j (Eq 3) of sequences S_i and S_j were calculated as correlation distances:

$$D_I(S_i, S_j) = cdist(M_i, M_j). \tag{4}$$

Participants

In total, 25 right-handed volunteers (19-34 years old, 10 female) gave informed, written consent for participation in the study after its nature had been explained to them. Participants reported no history of psychiatric or neurological disorders and no current use of any psychoactive medications. Three participants were excluded from the study because of excessive inter-scan movements (see *fMRI data acquisition and pre-processing*). The study was approved by the Cambridge Local Research Ethics Committee (Cambridge, UK).

Task

On each trial, participants saw sequence of items (oriented Gabor patches) displayed in the centre of the screen (Fig 1A). Each item was displayed on the screen for 2.4s (the whole four-item sequence 9.6s). Presentation of a sequence was followed by a delay of 4.8s during which only a fixation cue '+' was displayed on the screen. After the delay, participants either saw a response cue '*' in the centre of the screen indicating to manually recall the sequence exactly as they had just seen it, or a cue '–' indicating not to respond, and to wait for for the next sequence (rest phase; 10-18s). We used a four-button button-box where each button was mapped to a single item (see *Stimuli* below).

The recall cue appeared on 3/4 of the trials and the length of the recall period was limited to 7.2s. We omitted the recall phase for 1/4 of the trials to ensure a sufficient degree of de-correlation between the estimates of the BOLD signal for the delay and recall phases of the task. Each participant was presented with 72 trials (36 trials per scanning run) in addition to an initial practice session outside the scanner. In the practice session participants had to recall two individual sequences 16 times as they learned the mapping of items to button-box buttons. These sequences formed the learned sequences that could be compared with novel sequences. Participants were not informed that there were different types of trials.

Stimuli

All presented sequences were permutations of the same four items. The items were Gabor patches which only differed with respect to the orientation of the patch. Orientations of the patches were equally spaced (0, 45, 90, 135 degrees) to ensure all items were equally similar to

each other. The Gabor patches subtended a 6° visual angle around the fixation point in order to elicit an approximately foveal retinotopic representation. Stimuli were back-projected onto a screen in the scanner which participants viewed via a tilted mirror.

We used sequences of four items to ensure that the entire sequence would fall within the participants' short-term memory capacity and could be accurately retained in STM. If we had used longer sequences where participants might make errors (e.g. 8 items) then the representation of any given sequence would necessarily vary from trial to trial, and no consistent pattern of neural activity could be detected. All participants learned which four items corresponded to which buttons during a practice session before scanning. These mappings were shuffled between participants (8 different mappings) and controlled for heuristics (e.g. avoid buttons mapping orientations in a clockwise manner).

Over the course of the experiment we presented two learned sequences intermixed with novel sequences (previously unseen permutations of Gabors) so that in a 36-trial session participants recalled each novel sequence once and learned sequences twelve times each (Fig 1B). Over two scanning session this resulted in 48 trials with learned sequences and 24 trials with novel sequences.

Sequence generation

We chose the fourteen individual four-item sequences used in the experiment (2 learned, 12 novel) to maximise the predictive power of sequence representation models. We constrained the possible set of sequences with two criteria:

1. *Distinctiveness between all sequences*: to avoid the effects of interference, all sequences needed to be at least two edits apart in the Hamming distance space. For example, given a learned sequence $\{A, B, C, D\}$ we wanted to avoid a novel sequence $\{A, B, D, C\}$ as these share the two first items and hence the latter would only be a 'partially novel' sequence. Hence no novel sequences shared the first two items with the learned sequences.
2. *Distinctiveness between learned sequences*: the two learned sequences shared no items at common positions. This ensured that the representations of learned sequences would not interfere with each other and hence both could be learned to similar level of familiarity. Secondly, this increased the variance of the similarity scores between learned and novel sequences (see below).

Given these two constraints, we wanted to find a set of sequences which maximised two statistical power measures:

1. *Between-sequence similarity score entropy*: this was measured as the entropy of the lower triangle of the between-sequence similarity matrix (Fig 3B). The pairwise similarity matrix between 14 sequences has $14^2 = 196$ cells, but since it is diagonally identical only 84

cells can be used as a predictor of similarity for experimental data. Note that the maximum entropy set of scores would have an equal number of possible distances but since that is theoretically impossible we chose the closest distribution given the restrictions above (Fig 3D).

2. *Between-model dissimilarity*: defined as the correlation between pairwise similarity matrices (Eq 3) of different sequence representation models. The models were the Hamming distance model representing the positional encoding hypothesis (see *H1: Learning reduces noise in sequence representations*) and the n-gram models (see *H2: Learning recodes sequence representations*). We sought to maximise the dissimilarity between model predictions, that is, decrease the correlation between similarity matrices (Fig 3C).

The two measures described above, together with the constraints, were used as a cost function for a grid search over a space of all possible subsets of fourteen sequences ($k = 14$) out of possible total number of four-item sequences ($n =!4$). Since the Binomial coefficient of possible choices of sequences is ca 2×10^6 we used a Monte Carlo approach of randomly sampling 10^4 sets of sequences to get a distribution for cost function parameters. This processes resulted in a set of fourteen individual sequences which were used in the study: the properties of the sequence set are described on Fig 3. Importantly, by maximising the between-model dissimilarity we ensured that the correlation between the predictions made by the Hamming/positional model and the chunking/n-gram models were negative (Fig 3C).

To understand why the positional and chunking models make inversely correlated predictions, consider again the example given above: two sequences of same items FBI and BIF are similar from a 2-gram chunking perspective since both could be encoded using a 2-gram where B is followed by I (but share no items at common positions and are hence dissimilar in terms of item-position associations). Conversely, two sequences FBI and FIB share no item pairs (2-grams) and are hence dissimilar form a 2-gram chunking perspective but have both F at the first position and hence somewhat similar in terms of item-position model (Hamming distance).

fMRI data acquisition and pre-processing

Acquisition

Participants were scanned at the Medical Research Council Cognition and Brain Sciences Unit (Cambridge, UK) on a 3T Siemens Prisma MRI scanner using a 32-channel head coil and a simultaneous multi-slice data acquisition. Functional images were collected using 32 slices covering the whole brain (slice thickness 2 mm, in-plane resolution 2×2 mm) with acquisition time of 1.206 seconds, echo time of 30ms, and flip angle of 74 degrees. In addition, high-resolution MPRAGE structural images were acquired at 1mm isotropic resolution. (See

http://imaging.mrc-cbu.cam.ac.uk/imaging/ImagingSequences for detailed information.) Each participant performed two scanning runs and 510 scans were acquired per run. The initial ten volumes from the run were discarded to allow for T1 equilibration effects. Stimulus presentation was controlled by PsychToolbox software (Kleiner et al., 2007). The trials were rear projected onto a translucent screen outside the bore of the magnet and viewed via a mirror system attached to the head coil.

Anatomical data pre-processing

All fMRI data were pre-processed using *fMRIPprep* 1.1.7 (Esteban, Markiewicz, et al., 2018; Esteban, Blair, et al., 2018), which is based on *Nipype* 1.1.3 (Gorgolewski et al., 2011, 2018). The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using `N4BiasFieldCorrection` (Tustison et al., 2010, ANTs 2.2.0), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using `antsBrainExtraction.sh` (ANTs 2.2.0), using OASIS as target template. Brain surfaces were reconstructed using `recon-all` (Dale, Fischl, & Sereno, 1999, FreeSurfer 6.0.1), 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 grey-matter of Mindboggle (Klein et al., 2017). Spatial normalisation to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration with `antsRegistration` (Avants, Epstein, Grossman, & Gee, 2008, ANTs 2.2.0), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and grey-matter (GM) was performed on the brain-extracted T1w using `fast` (Zhang, Brady, & Smith, 2001, FSL 5.0.9).

Functional data pre-processing

The BOLD reference volume was co-registered to the T1w reference using `bbregister` (FreeSurfer) using boundary-based registration (Greve & Fischl, 2009). Co-registration 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) were estimated using `mcflirt` (Jenkinson, Bannister, Brady, & Smith, 2002, FSL 5.0.9). The BOLD time-series were slice-time corrected using `3dTshift` from AFNI (Cox & Hyde, 1997) package and then resampled onto their original, native space by applying a single, composite transform to correct for head motion and susceptibility distortions. Finally, the time-series were resampled to the MNI152 standard space (ICBM 152 Nonlinear Asymmetrical template version 2009c, Fonov et al., 2009) with a single interpolation step including head-motion transformation, susceptibility distortion correction,

and co-registrations to anatomical and template spaces. Volumetric resampling was performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimise the smoothing effects of other kernels (Lanczos, 1964). Surface resamplings were performed using `mri_vol2surf` (FreeSurfer).

Three participants were excluded from the study because more than 10% of the acquired volumes had extreme inter-scan movements (defined as inter-scan movement which exceeded a translation threshold of 0.5mm, rotation threshold of 1.33 degrees and between-images difference threshold of 0.035 calculated by dividing the summed squared difference of consecutive images by the squared global mean).

fMRI data analysis

To study sequence-based pattern similarity across all task phases we modelled the presentation, delay, and response phases of every trial (Fig 1A) as separate event regressors in the general linear model (GLM). We fitted a separate GLM for every event of interest by using an event-specific design matrix to obtain each event's estimate including a regressor for that event as well as another regressor for all other events (LS-S approach in Mumford, Turner, Ashby, & Poldrack, 2012). Besides event regressors, we added six head motion movement parameters and additional scan-specific noise regressors to the GLM (see *Functional data pre-processing* above). The regressors were convolved with the canonical hemodynamic response (as defined by SPM12 analysis package) and passed through a high-pass filter (128 seconds) to remove low-frequency noise. This process generated parameter estimates (beta-values) representing every trial's task phases for every voxel.

We segmented each participants' grey matter voxels into anatomically defined regions of interest (ROI, $n = 74$). These regions were specified by the Desikan-Killiany brain atlas (Desikan et al., 2006) and automatically identified and segmented for each participant using `mri_annotation2label` and `mri_label2vol` (FreeSurfer).

Representational similarity analysis

Representation of novel sequences

First, we created a representational dissimilarity matrix (RDM) $\mathbf{S}$ for the stimuli by calculating the pairwise distances s_{ij} between all 12 novel sequences $\{N_1, \ldots, N_{12}\}$:

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,12} \\ \vdots & \ddots & \vdots \\ s_{12,1} & \cdots & s_{12,12} \end{bmatrix},$$

$$s_{ij} = D(N_i, N_j)$$

where s_{ij} is the cell in the RDM $\mathbf{S}$ in row i and column j, and N_i and N_j are individual novel sequences. $D(N_i, N_j)$ is the distance measure corresponding to one of the four sequence representation models:

1. The positional model: Hamming distance (Eq 1)
2. The 2-item chunking model: n-gram distance with $n = 2$ (Eq 2)
3. The 3-item chunking model: n-gram distance with $n = 3$ (Eq 2)
4. The item mixture model : the item mixture distance (Eq 4)

Fig 3B,C show the RDMs for all of the model predictions and the correlation between the predictions. Next, we measured the pairwise distances between the voxel activity patterns of novel sequences:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \dots & a_{1,12} \\ \vdots & \ddots & \vdots \\ a_{12,1} & \dots & a_{12,12} \end{bmatrix}, \tag{5}$$

$$a_{ij} = cdist(P_i, P_j) = 1 - \frac{(P_i - \mu_{P_i}) \cdot (P_j - \mu_{P_j})}{||(P_i - \mu_{P_i})||_2 ||(P_j - \mu_{P_j})||_2} \tag{6}$$

where a_{ij} is the cell in the RDM $\mathbf{A}$ in row i and column j, and P_i and P_j are voxel activity patterns corresponding to novel sequences N_i and N_j. As shown by Eq 6, the pairwise voxel pattern dissimilarity is calculated as a correlation distance.

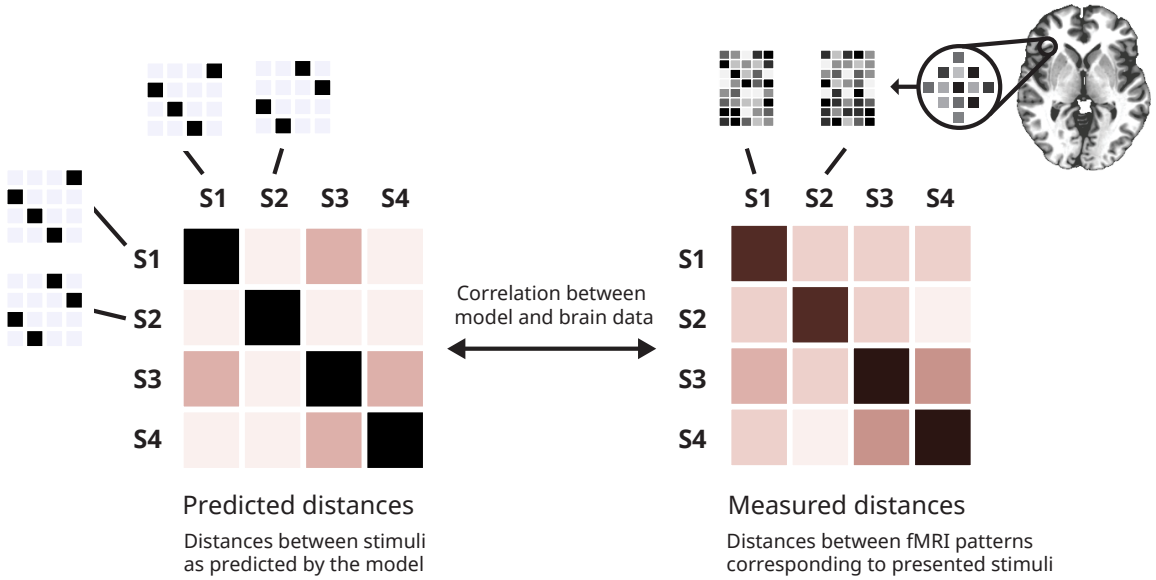
We then computed the Spearman's rank-order correlation between the stimulus and voxel pattern RDMs for every task phase p and ROI r:

$$r_{r,p} = \rho(\mathbf{rS}_{r,p}, \mathbf{rA}_{r,p}) = \frac{\mathbb{E}[(\mathbf{rS}_{r,p} - \mu_{\mathbf{rS}_{r,p}})(\mathbf{rA}_{r,p} - \mu_{\mathbf{rA}_{r,p}})]}{\sigma_{\mathbf{rS}_{r,p}} \sigma_{\mathbf{rA}_{r,p}}} \tag{7}$$

where ρ is the Pearson correlation coefficient applied to the ranks $\mathbf{rS}$ and $\mathbf{rA}$ of data $\mathbf{S}$ and $\mathbf{A}$ respectively.

Finally, to identify which ROIs represented novel sequences significantly above chance for all task phases we tested whether the Spearman correlation coefficients r were significantly positive across participants (see *Significance testing* below). The steps of the analysis are outlined on Fig 7.

Figure 7: Representational similarity analysis. Left: model prediction expressed as an RDM of pairwise between-stimulus distances (e.g. Hamming distance for the positional model). Right: measured RDM of voxel activity patterns elicited by the stimuli. The correlation between these two RDMs reflects the evidence for the representational model. The significance of the correlation can be evaluated via permuting the labels of the matrices and thus deriving the null-distribution (see *Significance testing*).



Noise ceiling estimation

Measurement noise in an fMRI experiment includes the physiological and neural noise in voxel activation patterns, fMRI measurement noise, and individual differences between subjects – even a perfect model would not result in a correlation of 1 with the voxel RDMs from each subject. Therefore an estimation of the noise ceiling is necessary to indicate how much variance in brain data – given the noise level – is expected to be explained by an ideal 'true' model.

We calculated the upper bound of the noise ceiling by finding the average correlation of each individual single-subject voxel RDM (Eq 5, 6) with the group mean, where the group mean serves as a surrogate for the perfect model. Because the individual distance structure is also averaged into this group mean, this value slightly overestimates the true ceiling. As a lower bound, each individual RDM was also correlated with the group mean in which this individual was removed. For more information about the noise ceiling please see Nili et al. (2014).

H1: Learning reduces noise in sequence representations

To test whether learning reduces noise in sequence representations we contrasted the noise levels in novel and learned sequence representations. Noise in sequence representations can be estimated by assuming that the voxel pattern similarity **A** is a function of the 'true' represen-

tational similarity between sequences $\mathbf{S}$ plus some measurement and/or cognitive noise:

$$\mathbf{A} = \mathbf{S} + \nu,$$

where ν is additive noise. In other words, here the noise is just the difference between predicted and measured similarity. Note that this is only a valid noise estimate when the predicted and measured similarity are significantly positively correlated (i.e. there is 'signal' in the channel).

If learning reduces noise in sequence representations then the noise in activity patterns generated by novel sequences ν_N should be greater than for learned sequences ν_L. To test this we measured whether the activity patterns of learned sequences were similar to novel sequences as predicted by the Hamming distance. The analysis followed exactly the same RSA steps as above, except instead of carrying it out within novel sequences we do this between novel and learned sequences. First, we computed the Hamming distances between individual learned and novel sequences $\mathbf{S}_{N,L}$, next the corresponding voxel pattern similarities $\mathbf{A}_{N,L}$ and finally computed the Spearman's rank-order correlation between the stimulus and voxel pattern RDMs exactly as above (Eq 7). If this measured correlation is significantly greater than the one within novel sequences ($r_{N,L} > r_N$) across participants, it follows that the noise level in learned representations is lower than in novel representations. This analysis was carried out for all task phases and in all ROIs where novel sequences were represented above chance, and finally tested for significance across participants.

The outcome of this analysis could fall into one of three categories:

1. No significant correlation: the probability of $r_{N,L}$ is less than the significance threshold ($p(r_{N,L}) < 10^{-4}$ see *Significance testing* below). This means that learned sequences are not represented positionally in this ROI and hence the test for noise levels is meaningless.
2. Significant correlation, but consistently smaller across participants than the within-novel sequences measure: $r_{N,L} < r_N$. Learned sequence representations are noisier than novel sequence representations.
3. Significant correlation, but consistently greater across participants than the within-novel sequences measure: $r_{N,L} > r_N$. Learned sequence representations are less noisy than novel sequence representations.

H2: Learning recodes sequence representations

We could not carry out RSA over the learned sequences similar to novel sequences as described above since we only used two individual learned sequences in the experiment (compared with 12 individual novel sequences). Due to limitations on the space of four-item sequences we could

not have have both 12 novel and learned sequences recalled by the participants. This is because there are only $4! = 24$ possible permutations of four items and hence the similarity between 12 novel and 12 learned sequences would be too great for to control for unwanted transfers of learning (i.e. learned and novel sequences sharing several items at same positions etc). *Sequence generation* in *Task* gives a detailed description of these restrictions.

However, each learned sequence was repeated 12 times in a single scanning run and 24 times over the course of the experiment. This allowed us to test whether an ROI encoded learned sequences by using support vector machine (SVM) classification on voxel patterns pertaining to either of the learned sequences. The details of the SVM testing, training, and significance assessment procedures are described below.

Linear classification of two learned sequences

We ran a binary classification of sequence identity (*learned sequence A* vs. *learned sequence B*) using only data from trials where a learned sequence was presented: 24 trials of *learned sequence A* and 24 trials of *learned sequence B*, resulting in 48 trials in total. For every ROI we labelled the voxel vectors comprising data from those trials according to their sequence identity and split the vectors into two data sets: a training set used to train a support vector machine (SVM, with a linear kernel and a regularization hyper-parameter $C = 1$) to assign correct labels to the activation patterns, and a test set (including one sample from each class) used to independently test the classification performance. The SVM classifier was trained to discriminate between the two individual sequences with the training data, and subsequently tested on the independent test data. The classification was performed with the LIBSVM (Chang & Lin, 2013) implementation. For every participant, the classification analysis resulted in a classification accuracy score for every ROI and task phase. The classification analysis was only carried out in the 9 ROIs where we found significant and sufficient (reaching the noise ceiling) evidence for novel sequence representation across task phases (Table 1). The mean accuracy score across the participants were then tested for group level significance.

Significance testing

We carried out the representational similarity and classification analyses for every task phase (encoding, delay, response; $n = 3$) and ROI ($n = 74$ for RSA, $n = 9$ for classification). To test whether the results were significantly different from chance across participants we used bootstrap sampling to create a null-distribution for every result and looked up the percentile for the observed result. We considered a result to be significant if it had a probability of $p < \alpha$ under the null distribution: this threshold α was derived by correcting an initial 5% threshold with the number of ROIs and task phases so that for RSA $\alpha = 0.05/74/3 \approx 10^{-4}$ and for

classification $\alpha = 0.05/9/3 \approx 10^{-3}$.

For RSA we shuffled the sequence labels randomly 100 times to compute 100 mean RSA correlation coefficients (Eq 7). For the classification analysis we permuted the correspondence between the test labels and data 100 different times to compute 100 mean classification accuracies for the testing labels. To this permuted distribution we added the score obtained with the correct labelling. We then obtained the distribution of group-level (across participants) mean scores by randomly sampling 1000 mean scores (with replacement) from each participant's permuted distribution. Next, we found the true group-level mean score's empirical probability based on its place in a rank ordering of this distribution.

Replication of analysis

The analysis scripts required to replicate the analysis of the fMRI data and all figures and tables presented in this paper are available at:
`https://gitlab.com/kristjankalm/fmri_seq_ltm/`.

The MRI data and participants' responses required to run the analyses are available in BIDS format at:
`https://www.mrc-cbu.cam.ac.uk/publications/opendata/`.

Acknowledgements

We would like to thank Jane Hall, Marta Correia, and Marius Mada for their assistance in setting up the experiments and collecting data. This research was supported by the Medical Research Council UK (SUAG/012/RG91365).

References

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26-41. doi: 10.1016/j.media.2007.06.004

Averbeck, B., Sohn, J., & Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature neuroscience*, *9*(2), 276–82. doi: 10.1038/nn1634

Berdyyeva, T., & Olson, C. (2010). Rank signals in four areas of macaque frontal cortex during selection of actions and objects in serial order. *Journal of Neurophysiology*, *104*(1), 141–59. doi: 10.1152/jn.00639.2009

Berdyyeva, T., & Olson, C. (2011). Relation of ordinal position signals to the expectation of reward and passage of time in four areas of the macaque frontal cortex. *Journal of Neurophysiology*, *105*(5), 2547–2559. doi: 10.1152/jn.00903.2010

Bettencourt, K. C., & Xu, Y. (2015). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience*, *19*(1), 150–157. doi: 10.1038/nn.4174

Brown, G. D., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*(1), 127.

Burgess, N., & Hitch, G. J. (1992). Toward a network model of the articulatory loop. *Journal of Memory and Language*, *31*(4), 429–460.

Buzsáki, G., & Tingley, D. (2018). Space and Time: The Hippocampus as a Sequence Generator. *Trends in Cognitive Sciences*, *22*(10), 853–869. doi: 10.1016/j.tics.2018.07.006

Carpenter, A. F., Georgopoulos, A. P., & Pellizzer, G. (1999). Motor Cortical Encoding of Serial Order in a Context-Recall Task. *Science*, *283*(5408), 1752–1757. doi: 10.1126/science.283.5408.1752

Chang, C.-c., & Lin, C.-j. (2013). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*, 1–39. doi: 10.1145/1961189.1961199

Christophel, T., Hebart, M., & Haynes, J. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, *32*(38), 12983–9. doi: 10.1523/jneurosci.0184-12.2012

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fmri data. *NMR in Biomedicine*, *10*(4-5), 171-178. doi: 10.1002/(sici)1099-1492(199706/08

Crowe, D., Zarco, W., Bartolo, R., & Merchant, H. (2014). Dynamic Representation of the Temporal and Sequential Structure of Rhythmic Movements in the Primate Medial Premotor Cortex. *Journal of Neuroscience*, *34*(36), 11972–11983. doi: 10.1523/jneurosci.2177-14.2014

Cumming, N., Page, M., & Norris, D. (2003). Testing a positional model of the hebb effect. *Memory*, *11*(1), 43–63.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179-194. doi: 10.1006/nimg.1998.0395

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, *88*(1), 2–19. doi: 10.1016/j.neuron.2015.09.019

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Devito, L., & Eichenbaum, H. (2011). Memory for the order of events in specific sequences: contributions of the hippocampus and medial prefrontal cortex. *The Journal of Neuroscience*, *31*(9), 3169–75. doi: 10.1523/jneurosci.4202-10.2011

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, *95*(23), 13988–13993. doi: 10.1073/pnas.95.23.13988

Du, Y., & Clark, J. E. (2017). New insights into statistical learning and chunk learning in implicit sequence acquisition. *Psychonomic Bulletin and Review*, *24*(4), 1225–1233. doi: 10.3758/s13423-016-1193-4

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. doi: 10.1038/nn.3428

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., ... Gorgolewski, K. J. (2018). Fmriprep 1.1.7. *Software*. doi: 10.5281/zenodo.852659

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., ... Gorgolewski, K. J. (2018). Fmriprep: a robust preprocessing pipeline for functional MRI. *bioRxiv*. doi: 10.1101/306951

Ester, E., Sprague, T., & Serences, J. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, *87*(4), 893–905. doi: 10.1016/j.neuron.2015.07.013

Fischer-Baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: Evidence from intrusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1426.

Fiser, J. (2009). Perceptual learning and representational learning in humans and animals. *Learning & behavior : a Psychonomic Society publication*, *37*(2), 141–53. doi: 10.3758/lb.37.2.141

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and

learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–30. doi: 10.1016/j.tics.2010.01.003

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47, Supplement 1*, S102. doi: 10.1016/s1053-8119(09)70884-5

Fortin, N., Agster, K., & Eichenbaum, H. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, *5*(5), 458–62. doi: 10.1038/nn834

Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., & Pine, J. (2001). Chunking mechanisms in Human Learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Gobet, F., Lloyd-Kelly, M., & Lane, P. C. R. (2016). What's in a Name? The Multiple Meanings of "Chunk" and "Chunking". *Frontiers in Psychology*, *7*(February), 1–5. doi: 10.3389/fpsyg.2016.00102

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. doi: 10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., ... Ghosh, S. (2018). Nipype. *Software*. doi: 10.5281/zenodo.596855

Graybiel, A. M., Aosaki, T., Flaherty, A. W., & Kimura, M. (1994). The basal ganglia and adaptive motor control. *Science*, *265*(5180), 1826–1831.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63-72. doi: 10.1016/j.neuroimage.2009.06.060

Hasselmo, M. E., Howard, M. W., Fotedar, M., & Datey, A. V. (2006). The Temporal Context Model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75–116.

Henson, R. (1998). Short-term memory for serial order: the Start-End Model. *Cognitive Psychology*, *36*(2), 73–137. doi: 10.1006/cogp.1998.0685

Henson, R. (1999). Positional information in short-term memory: relative or absolute? *Memory & Cognition*(27), 915–927.

Henson, R., & Burgess, N. (1997). Representations of Serial Order. *The Quarterly Journal of Experimental Psychology*(789284387).

Heusser, A. C., Poeppel, D., Ezzyat, Y., & Davachi, L. (2016). Episodic sequence memory is supported by a theta–gamma phase code. *Nature Neuroscience*, *19*(10), 1374–1380. doi: 10.1038/nn.4374

Hsieh, L.-T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal Activity Patterns Carry Information about Objects in Temporal Context. *Neuron*, *81*(5), 1165–1178. doi: 10.1016/j.neuron.2014.01.015

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, *140*(2), 339.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825-841. doi: 10.1006/nimg.2002.1132

Kahana, M. J., Mollison, M. V., & Addis, K. M. (2010). Positional cues in serial learning: The spin-list technique. *Memory & Cognition*, *38*(1), 92–101.

Kalm, K., & Norris, D. (2014). The representation of order information in auditory-verbal short-term memory. *The Journal of Neuroscience*, *34*, 6879–86. doi: 10.1523/jneurosci.4104-13.2014

Kesner, R. P., & Rolls, E. T. (2015). A computational theory of hippocampal function, and tests of the theory: New developments. *Neuroscience and Biobehavioral Reviews*, *48*, 92–147. doi: 10.1016/j.neubiorev.2014.11.009

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., ... Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. doi: 10.1371/journal.pcbi.1005350

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception 36 ECVP Abstract Supplement.*.

Kornysheva, K., & Diedrichsen, J. (2014). Human premotor areas parse sequences into their spatial and temporal features. *eLife*, *3*, e03043. doi: 10.7554/elife.03043

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76-85. doi: 10.1137/0701007

Lee, C., & Estes, W. (1981). Item and Order Information in Short-Term Memory: Evidence for Multilevel Perturbation Processes. *Journal of Experimental Psychology : Human Learning and Memory*, *7*(3), 149–169.

Legéndy, C. R. (2017). On the 'data stirring' role of the dentate gyrus of the hippocampus. *Reviews in the Neurosciences*, *28*(6), 599–615. doi: 10.1515/revneuro-2016-0080

Mayes, A. R., Isaac, C. L., Holdstock, J. S., Hunkin, N. M., Montaldi, D., Downes, J. J., ... Roberts, J. N. (2001). Memory for single items, word pairs, and temporal order of different kinds in a patient with selective hippocampal lesions. *Cognitive neuropsychology*, *18*(2), 97–123. doi: 10.1080/02643290125897

McClelland, J. L., McNaughton, B. L., & 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. *Psychological Review*, *102*(3), 419–457. doi: 10.1037/0033-295x.102.3.419

McKenzie, S., Keene, C. S., Farovik, A., Bladon, J., Place, R., Komorowski, R., & Eichenbaum,

H. (2016). Representation of memories in the cortical–hippocampal system: Results from the application of population similarity analyses. *Neurobiology of Learning and Memory*, *134*, 178–191. doi: 10.1016/j.nlm.2015.12.008

Merchant, H., Pérez, O., Zarco, W., & Gámez, J. (2013). Interval tuning in the primate medial premotor cortex as a general timing mechanism. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *33*(21), 9082–96. doi: 10.1523/jneurosci.5513-12.2013

Mumford, J. a., Turner, B. O., Ashby, F. G., & Poldrack, R. a. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–43. doi: 10.1016/j.neuroimage.2011.08.076

Murdock, B. (1997). *Context and mediators in a theory of distributed associative memory (TODAM2).* (Vol. 104). doi: 10.1037/0033-295x.104.4.839

Nieder, A. (2012). Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proceedings of the National Academy of Sciences*, *109*(29), 11860–11865. doi: 10.1073/pnas.1204580109

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). doi: 10.1371/journal.pcbi.1003553

Ninokura, Y., Mushiake, H., & Tanji, J. (2004). Integration of temporal order and object information in the monkey lateral prefrontal cortex. *J Neurophysiol*, *91*(1), 555–560. doi: 10.1152/jn.00694.2003

Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750. doi: 10.1073/pnas.0708424105

Pinsard, B., Boutin, A., Gabitov, E., Lungu, O., Benali, H., & Doyon, J. (2019). Consolidation alters motor sequence-specific distributed representations. *eLife*, *8*, 1–29. doi: 10.7554/elife.39324

Shimamura, A. P., Janowsky, J. S., & Squire, L. R. (1990). Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. *Neuropsychologia*, *28*(8), 803–813. doi: 10.1016/0028-3932(90)90004-8

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, 985–992. doi: 10.3115/1220175.1220299

Treves, A., & Rolls, E. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374–91. doi: 10.1002/hipo.450040319

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*,

29(6), 1310-1320. doi: 10.1109/tmi.2010.2046908

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, *22*(3), 477–483. doi: 10.1038/s41593-018-0335-6

Wiestler, T., & Diedrichsen, J. (2013). Skill learning strengthens cortical representations of motor sequences. *eLife*, *2*, 1–20. doi: 10.7554/elife.00801

Yokoi, A., Arbuckle, S. A., & Diedrichsen, J. (2018). The Role of Human Primary Motor Cortex in the Production of Skilled Finger Sequences. *The Journal of Neuroscience*, *38*(6), 1430–1442. doi: 10.1523/jneurosci.2798-17.2017

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45-57. doi: 10.1109/42.906424