Human sequence learning involves creating new neural representations not strengthening old 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.

December 14, 2018

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 neural representations by reducing noise or increasing signal gain. Alternatively, learning might cause the initial representation of the sequence 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 changes in neural 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 found no evidence for the SNR-change hypothesis. Instead, we observed that two brain regions in the dorsal visual processing stream encoded learned sequences as predicted by the chunking model. Our results suggest that learning-induced recoding elicits chunk-like representations of the learned sequence rather than simply strengthening the initial representations.

Introduction

Whenever we learn something new this must necessarily be supported by a change in the underlying neural representations. But what form does that change take? Research into neural learning mechanisms has shown that the signal-to-noise ratio of neural responses increases with repeated presentations of a stimulus (Dosher & Lu, 1998; Eldar, Cohen, & Niv, 2013; Gold, Bennett, & Sekuler, 1999; Schoups, Vogels, Qian, & Orban, 2001). Alternatively, repeated

presentations might elicit recoding of the stimulus using more efficient representations which reflect the changed statistics of the environment. For example, well-learned sequences may be recoded into chunks of successive items (Gobet et al., 2001; Gobet, Lloyd-Kelly, & Lane, 2016; Du & Clark, 2017). Usually this is assumed to be an incremental process where learning begins by forming low-level chunks consisting possibly of pairs of items, and then these chunks can be expanded or combined to form larger chunks. In some models this process is explicitly described as a way of developing more efficient representations (Robinet, Lemaire, & Gordon, 2011; Redlich, 1993). The crucial difference between these two learning approaches is that in the first case the codes for stimuli remain the same, whilst in the latter case a new model of the environment is inferred to recode stimuli.

Importantly, both of these learning mechanisms reduce the amount information required to represent the stimuli (Fiser, 2009; Fiser, Berkes, Orbán, & Lengyel, 2010) and therefore are hard to dissociate on the basis of behavioural measures such as improvement in recall or reaction times. However, in the context of sequence learning these mechanisms predict different neural response pattern similarities.

In this study we test which of these learning mechanisms is more likely given fMRI data from a task involving recall of both novel and learned sequences (Fig 1). First we define both learning mechanisms formally to predict the similarity structure between novel and learned sequences. We then compare the predicted similarity structure of each learning model to the actual similarity of neural activation patterns to infer which model is more likely given the fMRI data.

We found no evidence for the hypothesis that learning changes noise levels in initial sequence representations. Instead, we found that in two separate brain regions in the dorsal visual processing stream, novel sequence representations were recoded into chunk-like representations. We show that instead of simply making representations less noisy, learning led to a change in the form of the representations from simple item-position associations to chunks.

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.

Model of sequence representation

Novel sequences

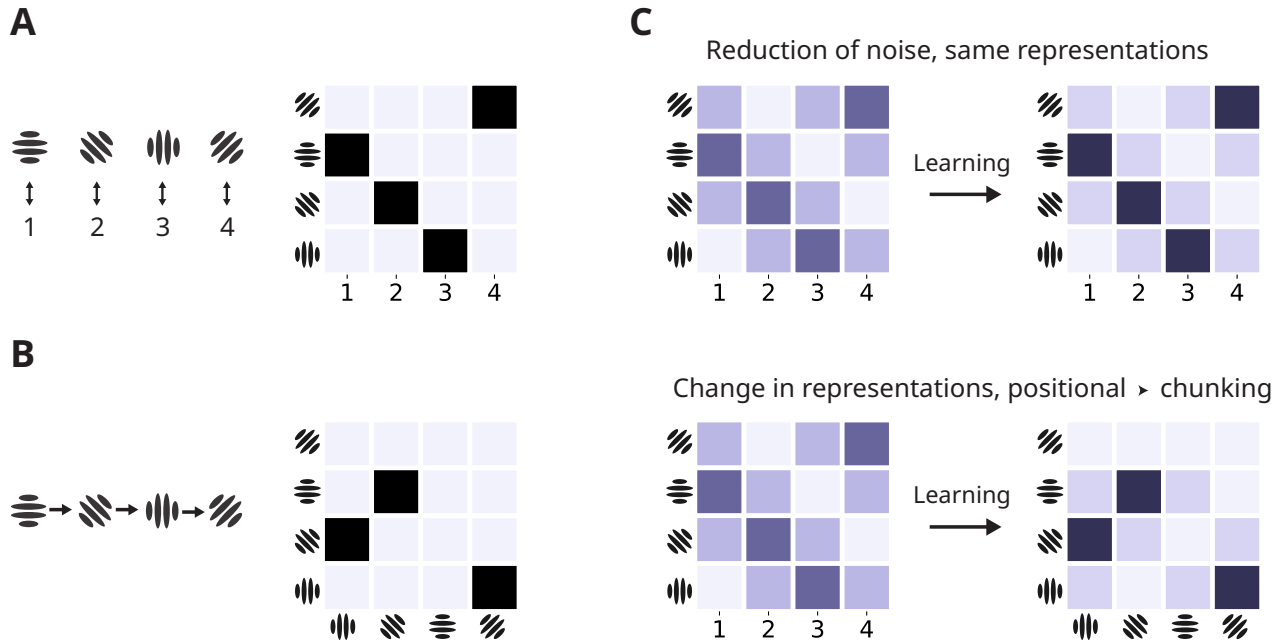
Previous research indicates that novel sequences are initially represented as position-item associations in the human brain (Heusser, Poeppel, Ezzyat, & Davachi, 2016; Hsieh, Gruber, Jenkins, & Ranganath, 2014; Kalm & Norris, 2014). Data from primates and rodents similarly suggest that sequences are initially encoded in terms of associations between items and their temporal positions (Fig 2A; Berdyyeva & Olson, 2010, 2011; Averbeck, Sohn, & Lee, 2006; Heusser et al., 2016; Ninokura, Mushiake, & Tanji, 2004).

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:

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

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

Figure 2: Sequence representation. 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.



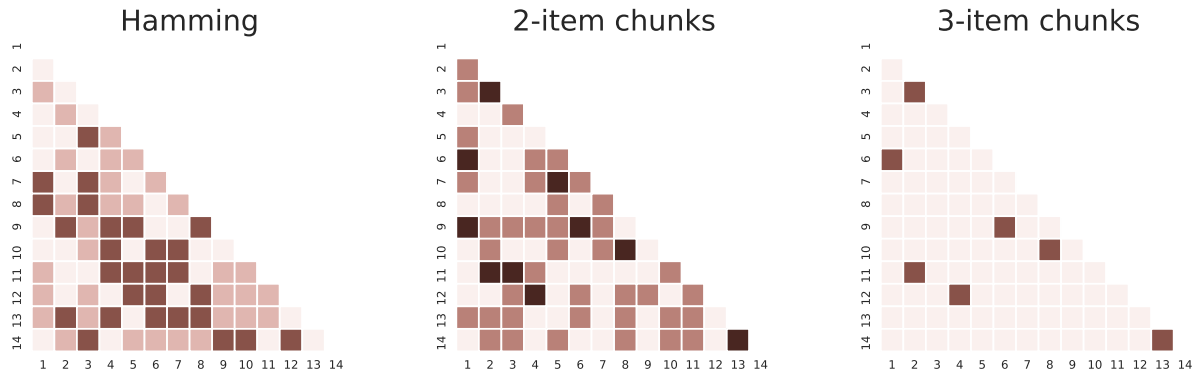
elicited by novel sequences should follow the Hamming distance (see *Representational similarity analysis* in *Methods* for details). Fig 3 (left panel) shows the similarity between all individual sequences used in this study's experiments according to the Hamming distance.

The model of novel sequence representation allows to test the first hypothesis: learning does not change representations but affects the signal-to-noise ratio. For this to be true neural activity elicited by both novel and learned sequences should be predicted by the Hamming model but the noise in novel sequences should be greater than learned sequences (see *H1: Learning reduces noise in sequence representations* in *Methods* for details). Alternatively, learning could elicit recoding of sequences from position-item associations to form more efficient representations.

Learned sequences

Previous research has shown that in many learning tasks sequences are recoded into chunks (Gobet et al., 2001, 2016; Du & Clark, 2017). 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

Figure 3: Between-sequence similarity matrices. Similarity of individual sequences presented in the experiments of this study as estimated by different sequence representation models.



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-sequences similarity is proportional to how many common chunks their share. For example, two chunked 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:

$$D_C(S_i, S_j) = |C_i \cap C_j| \tag{2}$$

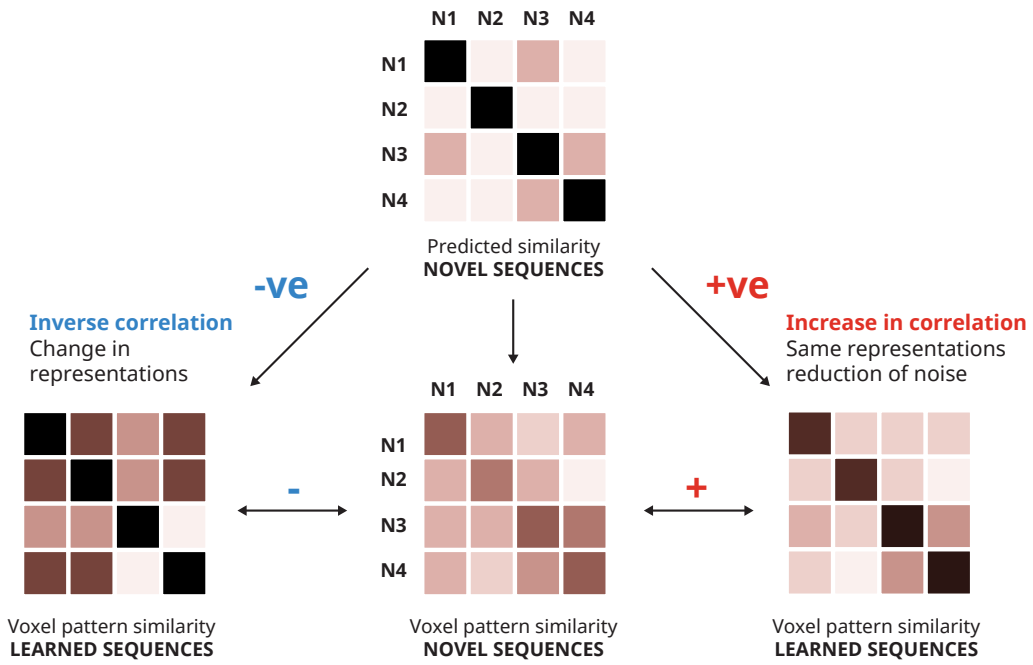
where C_i and C_j are the sets of n-grams required to encode sequences S_i and S_j respectively. Effectively, the chunking distance D_C counts the common members between two chunk sets.

The similarity 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 3 shows the similarity between individual sequences according to two chunking models: sequences represented as 2-grams and 3-grams.

Chunking models allow us to test the hypotheses that novel sequences are recoded during learning: we compare the neural activity similarity between novel and learned sequences to the predictions made by the positional and chunking models. Furthermore, we have chosen the

sequences in this study so that the predictions made by the positional (Hamming) and chunking model are inversely correlated to each other (see *Sequence generation* in *Methods* for details). This means that if novel sequences are recoded into chunks during the course of learning then the learned sequence representations should be also negatively correlated to the novel ones (see Fig 4 for an overview of hypothesis testing and *H2: Learning recodes sequence representations* in *Methods* for details).

Figure 4: Hypothesis testing with between-sequence similarity. First, we predict the between-sequence similarity for novel sequences according to the Hamming distance (top). Next, we correlate the prediction to the actual neural activity across the brain (middle, bottom). In brain regions where this correlation is significant, we test the two learning hypotheses: (H1) learning reduces noise in representations (right, bottom) or increase in Hamming similarity for learned sequences (compared to novel ones); or (H2) learning recodes novel sequences to chunks resulting in negative similarity between novel and learned sequences (left, bottom).



Results

Behavioural measures

In our task participants manually recalled a four-item sequence of Gabor patches after a 4.2s delay period (Fig 1). We calculated two behavioural measures of recall: how many items were recalled at correct positions, 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 average accuracy 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.

Evidence for learning models in fMRI data

We carried out the analysis in the dorsal visual processing stream which was further parcellated into 74 anatomically distinct regions of interest (ROI). In every one of these ROIs we first tested whether the region represented novel sequences, and if yes, then compared the representation of the learned sequences to the novel ones. This was done separately for all three task phases (presentation, delay, response).

H1: Learning reduces noise in sequence representations

If learned sequences are represented similarly to novel sequences but with less noise, then the activity patterns of learned sequences should be similar to novel sequences as predicted by the Hamming distance and this similarity should be on the average greater than within novel sequences (see *Representational similarity analysis* for the detailed derivation of the hypotheses). First, we found that a number of anatomical sub-regions in the dorsal visual processing stream represented novel sequences significantly above chance (Table 1).

However, we found no evidence for the noise reduction hypothesis in any of the ROIs for any three task phases (presentation, delay, response). Specifically, for learned sequences there was no significant correlation between the predictions made by the Hamming distance model and the voxel pattern similarity. In other words, we found no brain areas where learned sequences were also encoded positionally (like novel sequences) but with less noise.

Table 1: Evidence for the representation of novel sequences. Anatomical region prefixes indicate gyrus (G) or sulcus (S). For every task phase results are displayed separately for the left and right hemisphere. Values represent the probability of null hypothesis. Legend: *** $p < 10^{-5}$, ** $p < 10^{-4}$, * $p < 10^{-3}$.

	Presentation		*Delay*		*Response*	
Region	Left	Right	Left	Right	Left	Right
G Frontal Superior	**	***	***	*	***	***
G Precentral	***	***	**	***	***	*
S Central	**	**	**	***	**	**
G Paracentral	***	***	***	**	***	**
G Postcentral	***	**	***	**	**	**
S Postcentral	***	***	**	***	**	***
S Temporal Superior	**	***	***	***	***	***
G Parietal Superior	***	**	***	***	***	***
S Intraparietal Posterio-Transversal	**	***	***	*	**	***
G Parietal Inferior-Supramarginal	***	**	***	*	*	***
S Parieto-Occipital	***	***	***	*	**	*
G Precuneus	***	***	***	***	***	***
G Cuneus	***	*	**	**	**	***
G Occipital Middle		***	**	**	***	**

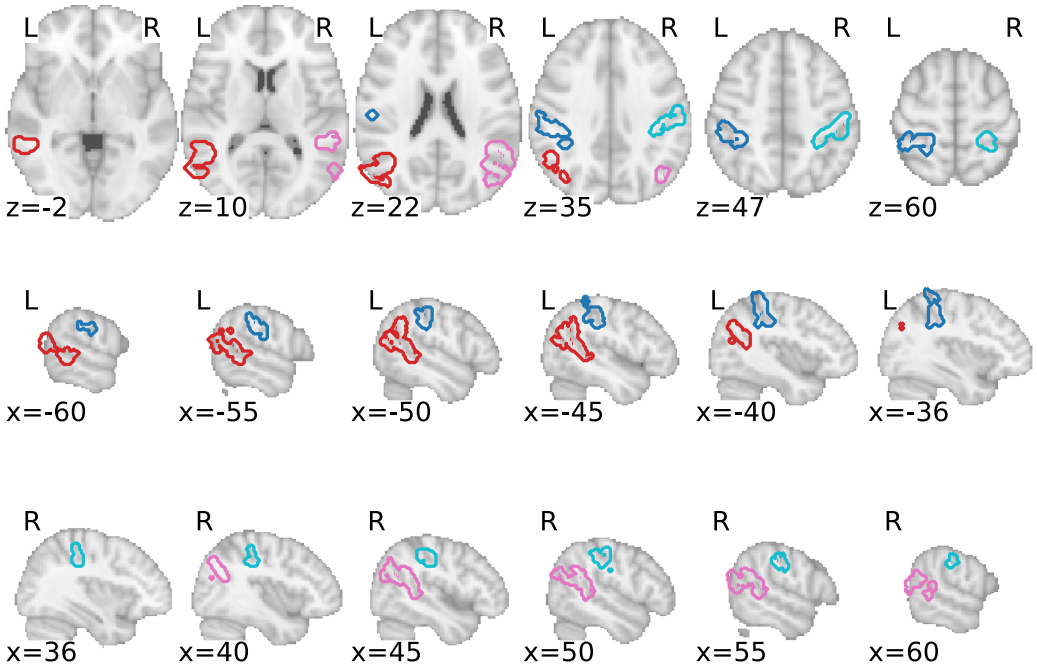
H2: Learning recodes sequences to chunks

If learned sequences come to be represented as chunks then the observed neural activity pattern similarity should reflect the predictions of the chunking model (see *H2: Learning recodes sequence representations* in *Methods*). Of the regions which encoded novel sequences (Table 1) we found significant evidence for the 2-gram chunking of learned sequences in two regions of interest: posterio-superior temporal sulcus and post-central sulcus (Fig 5, Table 2). The results were only significant in the response phase of the task (Table 2). No ROIs showed evidence for alternative chunking model based on 3-gram chunks.

Table 2: Evidence for the 2-gram chunking model. Anatomical region prefixes indicate gyrus (G) or sulcus (S). Values represent the probability of null hypothesis. Legend: *** $p < 10^{-5}$, ** $p < 10^{-4}$, * $p < 10^{-3}$.

	Presentation		*Delay*		*Response*	
Region	Left	Right	Left	Right	Left	Right
S Temporal Posterio-Superior					*	*
S Post-central						*

Figure 5: Anatomical localisation of learning effects. Regions of interest where learned sequence representations correlated with the 2-gram chunking model. Regions are shown for a single participant (P9) in the MNI152 standard space. Red: posterio-superior temporal sulcus; blue: post-central sulcus. Top: axial slices; middle: saggital slices, left hemisphere; bottom: saggital slices, right hemisphere.



Discussion

In the current study we contrasted two hypotheses about how neural representations change as a result of learning. First, we considered the possibility that repeated presentations might change the signal-to-noise ratio of existing neural 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 already know that sequences are initially represented as position-item associations (Heusser et al., 2016; Hsieh et al., 2014; Kalm & Norris, 2014; Berdyyeva & Olson, 2010). We were therefore able to test these contrasting predictions by comparing the patterns of activity elicited by novel and learned sequences with those predicted by position-item and chunking models.

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. We found no evidence that these brain regions also represented learned sequences positionally. Instead, we observed that two regions, which encoded novel sequences positionally, also encoded learned

sequences as predicted by the chunking model (Fig 5, Table 2). Our results indicate that learning-induced recoding elicits chunk-like representations of the learned sequence rather than simply strengthening the initial positional representation.

Of the two possible chunking mechanisms that were tested (2-gram and 3-gram chunking) we only found evidence for the 2-gram model. This result is not surprising if one considers how the predictions of the chunking models change as the size of the chunks increases. Here we only considered models with fixed chunk lengths: the 2-gram model only predicted similarity based on shared 2-grams and so forth. In other words, we excluded models with mixed chunk lengths (note that a four-item sequence could be encoded with a 2-gram and two 1-grams, etc.) since the resulting model space would be too large to test for meaningfully. However, the dissimilarity between chunked sequences increases monotonically with fixed chunk lengths. For example, when four-item sequences are encoded with a single chunk (4-gram) they are all maximally dissimilar to each other, thus reducing the between-sequence similarity to binary same-different evaluation. The same principle is evident when comparing the predictions made by the 2-gram and 3-gram models as displayed on Fig 3: the 3-gram model can only make meaningful predictions about seven sequence pairs out of possible 84. In other words, the 3-gram chunking model is too impoverished and statistically under-powered to offer a testable prediction or meaningful comparison to the 2-gram model.

Note that we are not suggesting a model in which 2-grams or 3-grams are the ultimate representations of sequences, merely that they may be involved in the development of more elaborate compressed representations. Previous research has suggested that chunking proceeds first by forming shorter chunks (e.g. pairs of items) and progressing incrementally to expand or combine these chunks until the whole sequence might be encoded as a single chunk (Gobet et al., 2001; Robinet et al., 2011). Similarly, Orban, Fiser, Aslin, and Lengyel (2008) have shown that visual scenes are encoded into higher order chunks beyond pairwise item-item associations. Our results agree with these findings and indicate that compositionally learned sequences are significantly more similar to chunks than item-position associations. The exact nature and formation of the chunked representation requires further research (but see Kikumoto & Mayr, 2018).

Finally, we suggest that the change from positional to chunked representations is here mediated both by our experimental task and from an optimal learning perspective. Re-coding sequences from position-item associations into chunks is also a superior strategy compared to strengthening of the position-item associations when dealing with multiple overlapping sequences. 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 item-position associations would result in significant interference between the learned associations. In other words, learning overlapping sequences increases variance in association weights and makes such representation less efficient with every new to-be-learned sequence.

In sum our results suggest that sequence learning results in chunk-like representations of sequences. We suggest that this learning mechanism has arisen from the limited capacity of human STM and that learning sequences via strengthening position-item associations is suboptimal compared to chunking.

Methods

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 practise 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 both novel sequences 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. Second, 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 (Hamming distance, Fig 3, left). The pairwise similarity matrix between 14 sequences has $14^2 = 196$ cells, but since it it 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 Hamming distances but since that is theoretically impossible we chose the closest distribution given the restrictions above (Fig 6A).

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 6B).

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 6 below. 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 6B).

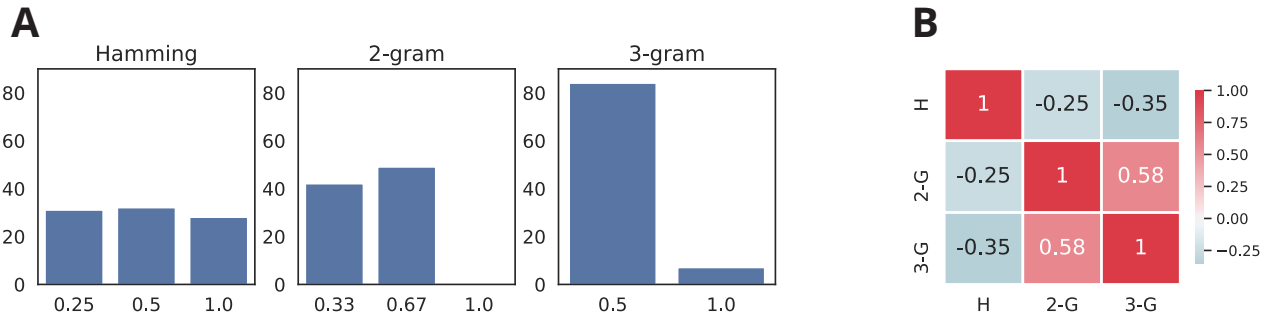
To understand why the positional and chunking models make inversely correlated predictions, consider again the example given above: two sequences of same items FBI and BIF are similar from a 2-gram chunking perspective since both could be encoded using a 2-gram where B is followed by I (but share no 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

Figure 6: Sequences used in the task. (A) Distribution of similarity scores between all sequences used in the study. Left: Hamming distance, middle: 2-gram chunks, right: 3-gram chunks. (B) Between-model similarity measured as Pearson's correlation coefficient between model produced pair-wise similarity matrices for the fourteen individual sequences. H: Hamming, 2-G: 2-gram chunks, 3-G: 3-gram chunks.



volumes from the run were discarded to allow for T1 equilibriation 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 (K. Gorgolewski et al., 2011; K. J. Gorgolewski et al., 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 using `mri_annotation2label` and `mri_label2vol` (FreeSurfer). In every ROI we carried out three analyses to detect the following effects: (1) representation of novel sequences; (2) H1 : learning reduces noise in sequence representations; and (3) H2: learning recodes sequence representations.

Representational similarity analysis

Novel sequences

First, we created a pairwise between-stimulus similarity matrix $\mathbf{S}$ by calculating the pairwise similarity s_{ij} between all 12 novel sequences $\{N_1, \ldots, N_{12}\}$ using the Hamming distance:

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

$$s_{ij} = D_H(N_i, N_j) \tag{3}$$

where s_{ij} is the cell in the similarity matrix $\mathbf{S}$ in row i and column j, and N_i and N_j are individual novel sequences. The Hamming distance between two sequences of equal length k measures how many items were not retained at their original positions:

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \tag{4}$$

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 N_i and N_j respectively. The Hamming distance therefore assumes that sequences are represented as item-position mappings.

Next, we measured the similarity between the neural activity patterns as a pairwise similarity matrix $\mathbf{A}$ for the same novel sequences:

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

$$a_{ij} = corr(P_i, P_j) = \frac{\mathbb{E}[(P_i - \mu_{P_i})(P_j - \mu_{P_j})]}{\sigma_{P_i}\sigma_{P_j}} \tag{5}$$

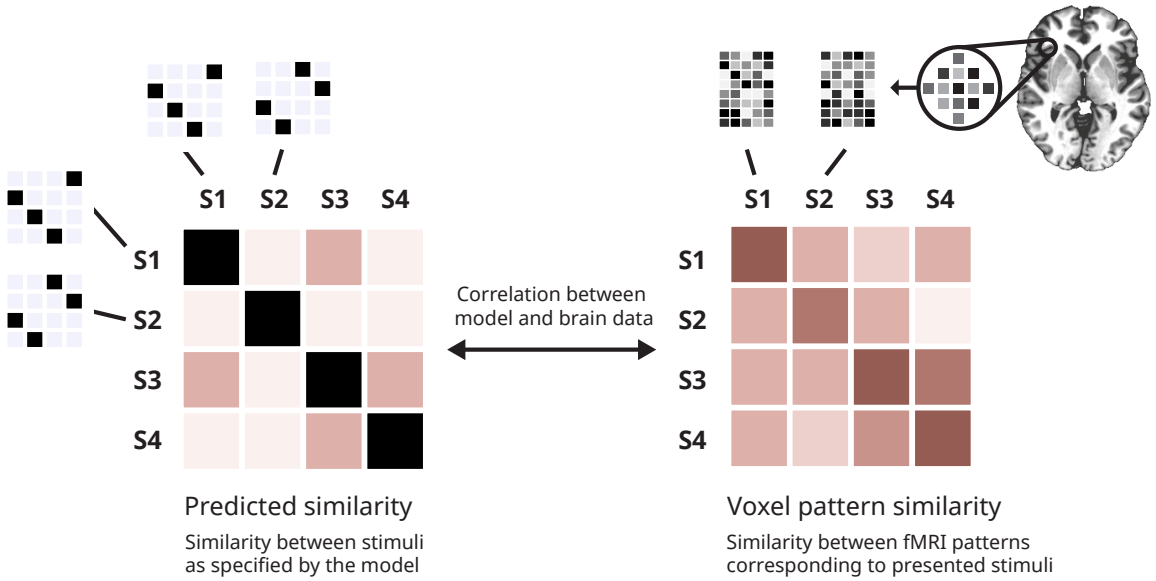
where a_{ij} is the cell in the similarity matrix $\mathbf{A}$ in row i and column j, and P_i and P_j are voxel activity patterns corresponding to novel sequences N_i and N_j. As shown by Eq 5, the between-pattern similarity is measured as Pearson's correlation coefficient.

We then computed the correlation between the stimulus and pattern similarity matrices for every task phase p and ROI r:

$$\rho_{r,p} = corr(\mathbf{S}_{r,p}, \mathbf{A}_{r,p}) = \frac{\mathbb{E}[(\mathbf{S}_{r,p} - \mu_{\mathbf{S}_{r,p}})(\mathbf{A}_{r,p} - \mu_{\mathbf{A}_{r,p}})]}{\sigma_{\mathbf{S}_{r,p}} \sigma_{\mathbf{A}_{r,p}}} \tag{6}$$

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

Figure 7: Representational similarity analysis. Left: predicted similarity between stimuli as evaluated by a distance function (e.g. Hamming distance). Right: measured fMRI pattern similarity between activity patterns elicited by the stimuli. The correlation between these two representational similarity matrices reflects the evidence for the representational model implemented by the distance function. The significance of the correlation can be evaluated via permuting the labels of the matrices and thus deriving the null-distribution (see *Significance testing*).



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 $\mathbf{A}$ is a function of the 'true' represen-

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

$$\mathbf{A} = \mathbf{S} + \nu, \tag{7}$$

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 correlated the predicted similarity matrix with the voxel activity pattern similarity matrix:

$$\rho_{N,L} = corr(\mathbf{S}_{N,L}, \mathbf{A}_{N,L}) = \frac{\mathbb{E}[(\mathbf{S}_{N,L} - \mu_{\mathbf{S}_{N,L}})(\mathbf{A}_{N,L} - \mu_{\mathbf{A}_{N,L}})]}{\sigma_{\mathbf{S}_{N,L}} \sigma_{\mathbf{A}_{N,L}}}. \tag{8}$$

Specifically, if this similarity is greater than the one within novel sequences ($\rho_{N,L} > \rho_N$) 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 $\rho_{N,L}$ is less than the significance threshold ($p(\rho_{N,L}) < 10^{-3}$ 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: $\rho_{N,L} < \rho_N$. Learned sequence representations are noisier than novel sequence representations.
3. Significant correlation, but consistently greater across participants than the within-novel sequences measure: $\rho_{N,L} > \rho_N$. Learned sequence representations are less noisy than novel sequence representations.

H2: Learning recodes sequence representations

Here we define chunks as n-grams of consecutive sequence items: for example, a four-item sequence $\{A, B, C, D\}$ can be represented by three 2-grams AB, BC, and CD. We define

the similarity between chunked sequences as the number of common chunks they share. This allows us to derive 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. This was further normalised to make the pairwise distance proportional to the total number of possible n-grams in the sequence. For example, for 2-gram distance between two sequences is:

$$D_{2-gram} = \sum_{i=1}^{k-1} | \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} \neq \begin{pmatrix} y_i \\ y_{i+1} \end{pmatrix} | \tag{9}$$

where x_i and y_i are the i-th items from sequences S_i and S_j respectively. We calculated pairwise between-sequence chunking similarity for both 2-gram and 3-gram chunking.

The similarity prediction made by the chunking distance D_C is fundamentally different from the prediction made by the Hamming distance D_H (Eq 4): the chunking distance assumes that sequences are encoded as item-item associations whilst the Hamming distance assumes that sequences are encoded as item-position associations. To test whether learned sequences were represented as chunks we ran the RSA between novel and learned sequences exactly as described above (Eq 8), except using the chunking distance (Eq 9) instead of the Hamming distance. 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.

Importantly, we chose the sequences in the experiment so that the similarity predictions made by the positional model for novel sequences and the chunking model are inversely correlated with each other (see *Sequence generation* and Fig 6 for details). In other words, when we construct two pairwise between-stimulus similarity matrices for the sequences used in this study: one with the Hamming distance and one with 2-gram distance (Fig 3, middle vs right) then the correlation between the two matrices is negative:

$$\rho_{Hamming,2-gram} = corr(\mathbf{S}_{Hamming}, \mathbf{S}_{2-gram}) = -0.25.$$

See Fig 6 for a full correlation matrix between models. The negative correlation between two predictions allow us to test for both learning hypotheses in a single step: a significant increase in Hamming similarity for learned sequences (compared to novel ones) is evidence for H1 (learning reduces noise) and significant negative correlation between novel and learned sequences is evidence for H2 (learning recodes novel sequences to chunks, see Fig 4 for a graphical

overview).

Significance testing

We carried out the representational similarity analyses for every task phase (encoding, delay, response; $n = 3$) and ROI ($n = 74$). To test whether the RSA 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 $p < 10^{-4}$ 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 $\alpha = 0.05/74/3 \approx 10^{-4}$).

First, we shuffled the sequence labels randomly 100 times to compute 100 mean RSA correlation coefficients (Eq 6 and Eq 8). To this permuted distribution of coefficients 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. The peak percentiles of significance ($p < 10^{-4}$) are limited by the number of samples producing the randomised probability distribution at the group level.

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).

Competing Interests

There are no competing interests to report.

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

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

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

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

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

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

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

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, *402*(6758), 176–178. doi: 10.1038/46027

Gorgolewski, K., 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

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

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

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

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

Kikumoto, A., & Mayr, U. (2018). Decoding hierarchical control of sequential behavior in oscillatory EEG activity. *eLife*, *7*, 1–36. doi: 10.7554/eLife.38550

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.*.

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

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

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

Redlich, A. N. (1993). Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Computation*, *5*(2), 289–304. doi: 10.1162/neco.1993.5.2.289

Robinet, V., Lemaire, B., & Gordon, M. B. (2011, aug). MDLChunker: A MDL-Based Cognitive Model of Inductive Learning. *Cognitive Science*, *35*, 1352–1389. doi: 10.1111/j.1551-6709.2011.01188.x

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*(6846), 549–553. doi: 10.1038/35087601

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

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