Sequence learning recodes 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.

Abstract

Here we contrast two computational models of sequence learning. The associative learner posits that learning proceeds by strengthening existing association weights. Alternatively, chunk learning posits that learning creates new and more efficient representations of the learned sequences. Importantly, both models propose that humans act as optimal learners but capture different statistics of the stimuli in their internal model. Furthermore, these models make dissociable predictions as to how learning changes the representation of sequences. We tested these predictions by using fMRI to extract neural activity patters from the dorsal visual processing stream during a sequence recall task. We observed that only the chunking account can explain the similarity of neural activity patterns. Specifically, the encoding of sequences in the posterior parietal cortex changed from representing novel sequences as individual items to representing them as chunks after they had been presented several times. Furthermore, we show that associative learning without recoding is not theoretically capable of supporting long-term storage of multiple overlapping sequences.

Introduction

Here we investigate the neural mechanism of learning short visual sequences. The ability to remember or to perform events or actions in the correct order is critical to the performance of almost all cognitive tasks [1]. Understanding human sequence learning mechanism is crucial not only for understanding normal cognition, but also to understand the nature of the impairments and disabilities that follow when sequence learning is disrupted [2, 3].

In this study we ask whether the changes in neural activity during sequence learning reflect an optimal learning strategy. An optimal learner is an agent whose internal model reflects the statistics of the environment [4, 5], and human learning has been shown to follow the optimal model in a wide range of domains such as speech and language [6, 7], visual scenes and objects [8, 9, 10, 11, 12], and sensorimotor control [13, 14].

However, statistical regularities across sequences can be represented in multiple ways [1].

First, sequences can be represented as simple associations (Fig 1A-B) and their statistics represented by weighting the associations based on their relative frequency (Fig 1C-D). An ideal learner would update the association weights as new data comes in to reflect the statistics of the environment. Alternatively, learning can proceed by recoding frequently occurring associations using new latent representations. The latter approach has been termed 'chunking' in cognitive literature [15, 16] to describe learning where complex objects (words, faces) are constructed from lower-level features (phonemes, syllables, oriented lines).

Recent research on sequence learning has provided evidence for both accounts across multiple time scales and modalities. Manual motor skill learning has been shown to decrease noise in learned representations [17, 18, 19] whilst not changing the representations of individual items in the sequence [20, 21]. Similarly, in the auditory domain frequently co-occurring sequence items elicit a neural response that indicates an increase in association strength [22]. Contrastingly, chunking has been observed widely in tasks where separate movements are integrated into a unified sequence [23], and in auditory-verbal sequence learning [24, 25, 26] where multiple co-occurring sequence items are bound together and recalled in all-or-nothing fashion [27, 28]. 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.

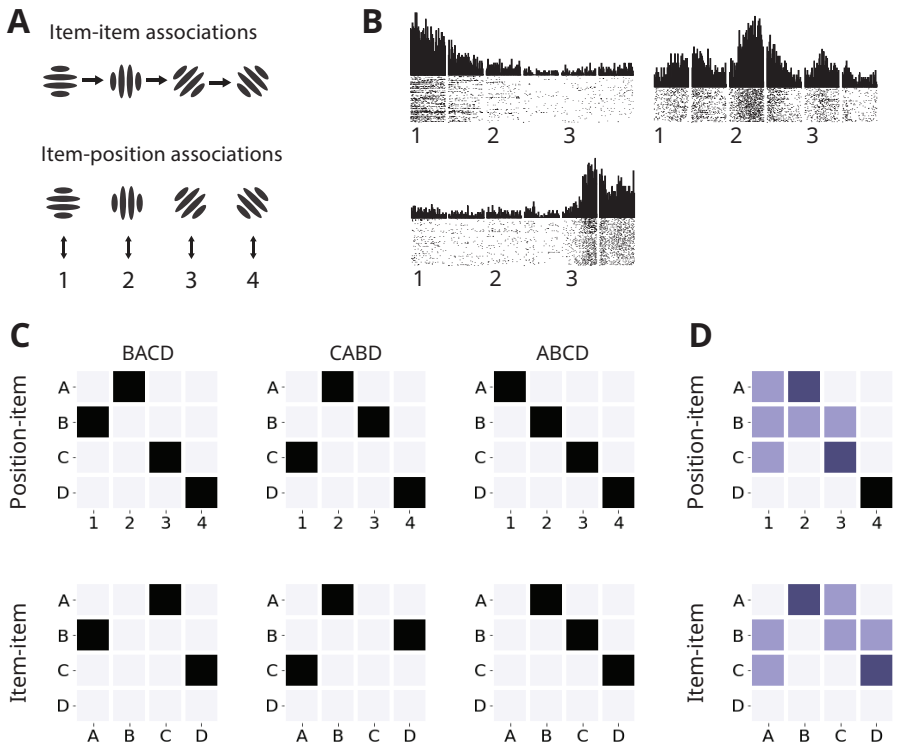
However, no sequence learning studies have directly compared the predictions of the two accounts: this requires formulating both mechanisms explicitly as computational models and testing their predictions with neural data. In this study we formally derive the associative and chunk recoding models in the context of Bayesian optimal learner. We show that the two accounts make dissociable predictions as to how individual items are encoded, and these predictions can be expressed in terms of the similarity of neural pattern activity. We used fMRI to extract neural activity patterns of individual sequences from the dorsal visual processing stream during a sequence recall task. We observed that only the recoding account can explain the similarity of neural activity patterns. Specifically, the encoding of sequences in the posterior parietal cortex changed from representing novel sequences as individual items to representing them as chunks after they had been presented several times. In other words, the representation

of the sequences changed to encode the evolving statistical regularities in the input.

Furthermore, we demonstrate that an associative learner is theoretically incapable of learning multiple overlapping sequences such as the ones presented in this study. We also show that latent codes (or chunks) extracted from sequences cannot share the same representational space with associative representations.

Taken together our findings represent strong theoretical and empirical evidence for a specific learning mechanism: human learning of short visual sequences proceeds by recoding initial sequence representations with new ones within the same brain regions. Such recoding is necessary to enable efficient behaviour in complex cognitive tasks.

Fig 1: Sequence representation. (A) Sequences are associative codes which are formed either by associating successive items to each other (item-item associations) or by associating items to some external signal specifying the temporal context for sequences (item-position associations). (B) Item-position associations in monkey prefrontal cortex as observed by Berdyyeva and Olson [29]. Each subplot displays spiking activity for a particular neuron: the first one responds most to items at the beginning of a three-item sequence, the second for the ones in the middle, and the last one for items at the end of the sequence. Numbers on x-axis mark the onset of the stimulus events. (C) Visual representation of three sequences as position-item associations matrix (top row) or item-item associations matrix (bottom row). (D) Matrices of associative weights expressing the statistics across all three sequences.

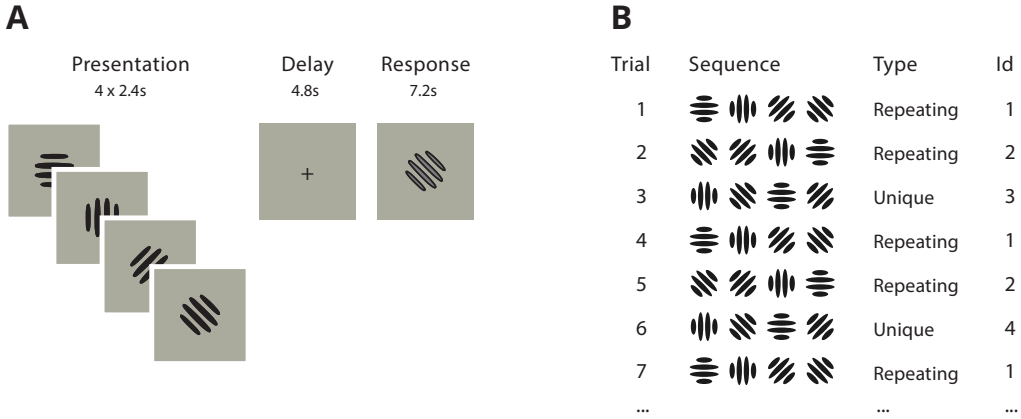


Results

Behaviour

We used a task requiring ordered recall of a sequence of simple visual stimuli, where some of the sequences are presented only once (*unique* sequences) and some are presented multiple times (*repeating* sequences, Fig 2). Only two individual sequences were repeated and we first presented them 12 times each during a practise session. This ensured that those two individual sequences were learned to criterion before the beginning of the main experiment. We proceeded to present the two familiar repeating sequences interleaved with novel unique sequences (Fig 2B).

Fig 2: Task. (A) Single trial: participants had to recall a sequence of four Gabor patches in the order they were presented after a short delay. The size of the stimuli within the display area is exaggerated for illustrative purposes. (B) Trial types and progression: 2/3 of the trials were repetitions of the same two individual sequences (*repeated* sequences), while for the 1/3 of the trials were novel unseen orderings of the items (*unique* sequences). Repeating and unique sequences were chosen for each participant so that they shared no statistical regularities across individual item positions (see *Methods* for details).



At every trial participants had to recall the presented sequence after a 4.8s delay period using a button-box. 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 unique and repeating sequences: 0.96 vs. 0.97, with no significant difference across subjects ($p = 0.22, df = 21$). This was expected since both unique and repeating sequences were only four items long

and should therefore fit within participants' short term memory span. However, participants were consistently faster in recalling repeating sequences: the average time between consecutive key presses was 0.018 seconds shorter for repeating sequences ($t = -3.04, p = 0.007, df = 21$). This shows that unique and repeating sequences were processed differently by participants.

Neural representation of sequences

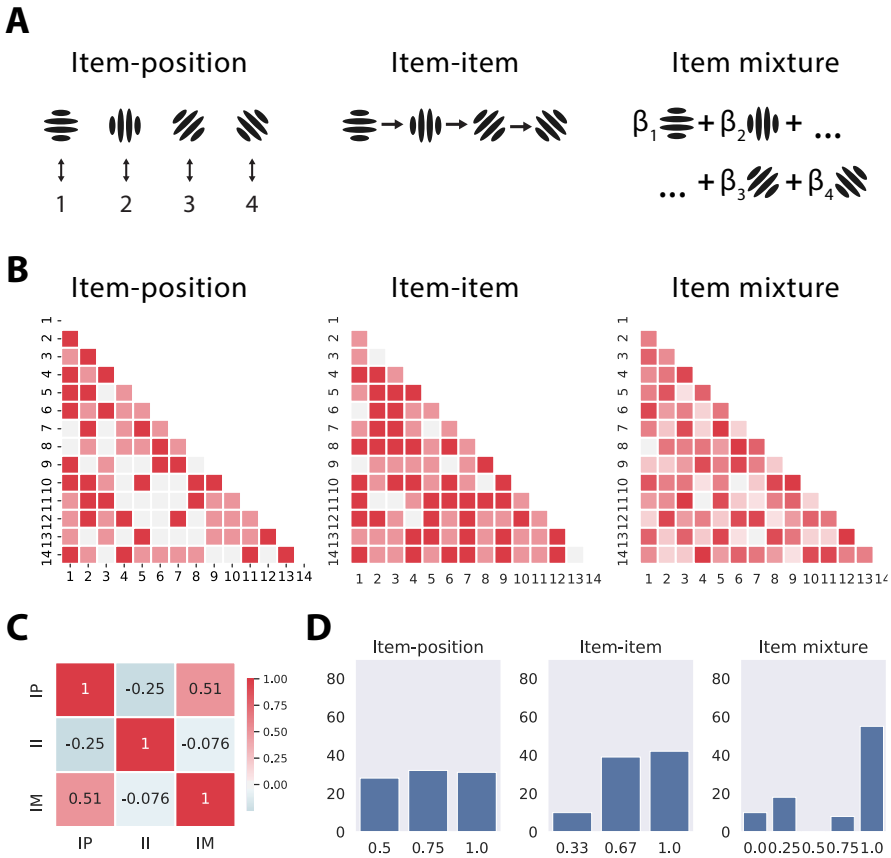
First, we sought to identify which brain regions are involved in sequence representation as the effects of learning can only be established in regions where such representations have been observed. For this purpose we used fMRI activation data only from trials of unique sequences – as there were no shared regularities between the unique and repeated sequences the representation of non-repeated unique sequences should not be affected by learning of the repeating ones.

We defined two formal sequence representation models to be sensitive to both types of neural sequence representations: past research has shown that sequences are represented in rodent, primate, and humans brains as both item-position [30, 31, 32] and item-item associations [33] depending on task type and anatomical area (Fig 1, also see [1] for a review). We also defined an additional control model which tested for a null-hypothesis that instead of sequence representations neural activity could be better explained by the superposition of patterns for constituent individual items in the sequence, called the *item mixture model* (see Fig 3 and Yokoi et al. [21]). The full formal description of sequence representation models and their parameters can be found in *Sequence representation models for RSA* in *Methods*.

We parcellated the dorsal visual processing stream bilaterally into 74 anatomically distinct regions of interest (ROI). In every ROI we fit the sequence representation models to voxel activation patterns corresponding to unique sequences. The model fits showed how well each model predicted the distance between pairs of voxel activity patterns corresponding to individual sequences (see *Representational similarity analysis (RSA)* in *Methods*). This was done separately for all three task phases (presentation, delay, response; Fig 2).

Besides identifying the best-fitting model we determined the noise ceiling for the model fits in each of the ROIs to determine whether any of the models also accounted for the representational

Fig 3: Models of neural sequence representation. (A) Visualisation of how models represent the order of items. β_i represent mixture weights for individual items i (see *Sequence representation models for RSA* in *Methods* for details). (B) Distance between the 14 individual sequences presented in the task of this study as estimated by different 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). IP: Item-position, II: Item-item, IM: Item mixture. (D) Distribution of normalised pairwise similarity distance values per model. The spread of the distance values indicates the predictive power of the model.



structure observed across participants. The noise ceiling is the expected fit with data achieved by the theoretical true model given the noise (see *Noise 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 item-position model and the item mixture model reached the noise ceiling

We found that only the item-position model and the item mixture model reached the noise ceiling for unique sequence representation. This was so in 14 of the 74 ROIs bilaterally (Table

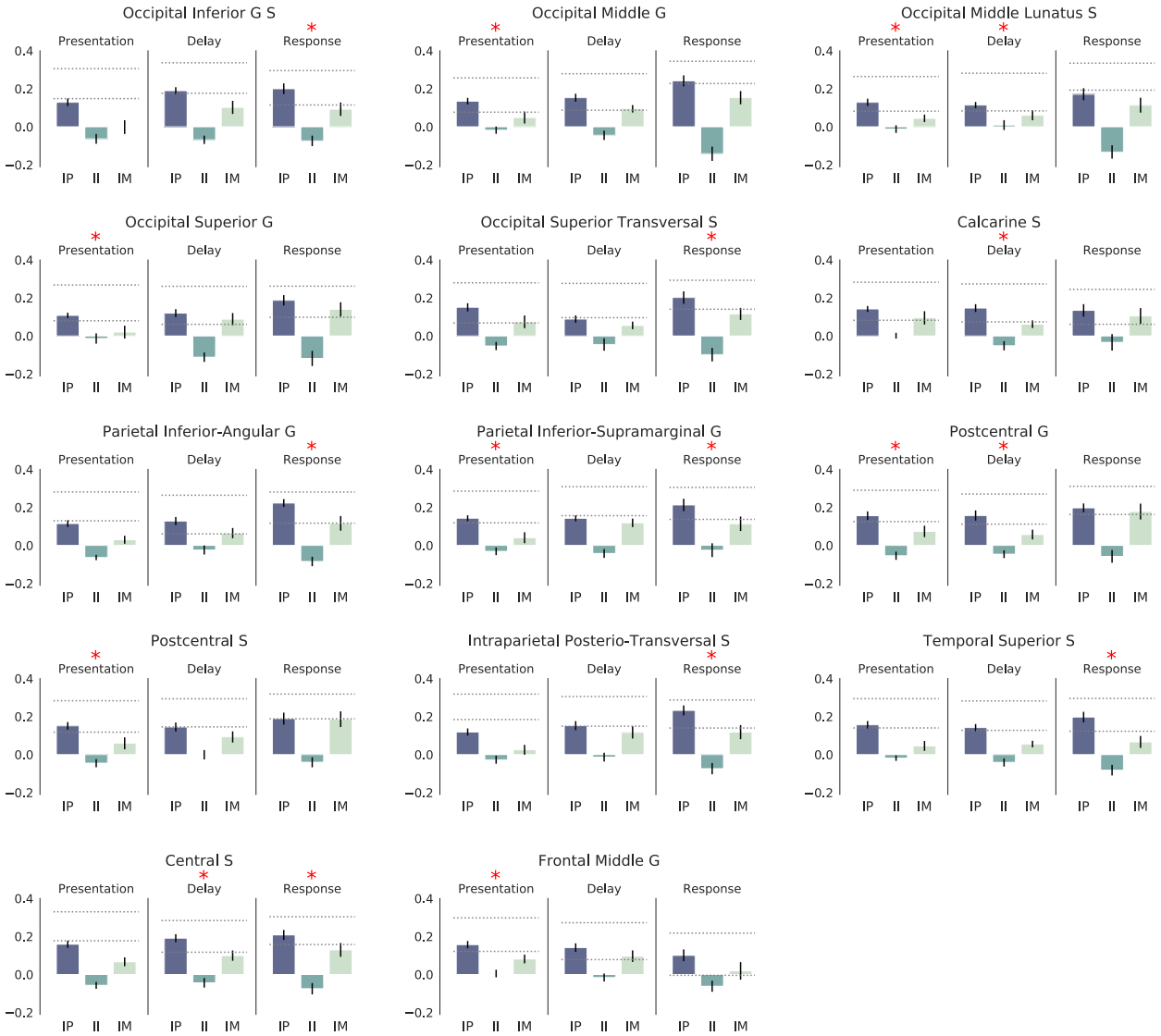
1). No item-item model reached the lower bounds of the noise ceiling in any ROI. This suggests that the item-item models are not a good fit for sequence representations anywhere in the dorsal visual processing stream. Since the predictions of the positional and item-item models are inversely correlated, such relationship between the model fits was expected: both cannot be positively correlated with the same data (see Fig 3C and *Sequence representation models for RSA* in *Methods*).

Table 1: Evidence for the representation of unique sequences. Anatomical region suffixes indicate gyrus (G) or sulcus (S). Asterisks ($*$) represent the item-position model and daggers ($\dagger$) the item-item model reaching the lower bound of the noise ceiling. The lower noise ceilings were significantly greater than zero for all regions displayed in the table ($df = 21, p < 10^{-3}$). Regions in which the item mixture model reached the noise ceiling in any of the task phases are displayed in italics.

Lobe	Name	Presentation	Delay	Recall
Frontal	Central S		$*$	$*$
Frontal	*Frontal Middle G*	$*$	$*,\dagger$	$\dagger$
Occipital	*Calcarine S*	$*,\dagger$	$*$	$\dagger$
Occipital	Occipital Inferior G S			$*$
Occipital	*Occipital Middle G*	$*$	$*,\dagger$	
Occipital	Occipital Middle Lunatus S	$*$		
Occipital	*Occipital Superior G*	$*$	$\dagger$	$*,\dagger$
Occipital	*Occipital Superior Transversal S*	$*,\dagger$		$*$
Parietal	Intraparietal Posterio-Transversal S			$*$
Parietal	*Parietal Inferior-Angular G*		$*,\dagger$	$*$
Parietal	Parietal Inferior-Supramarginal G	$*$		
Parietal	Postcentral G	$*$	$*$	
Parietal	Postcentral S	$*$		
Temporal	Temporal Superior S			$*$

The item mixture model reached the noise ceiling in several visual, motor, and parietal regions for all task phases (Fig 4, Table 1). The item mixture 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 [21] we observed that the item mixture 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 mixture model did not reach the noise ceiling

Fig 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 and task phases where the fit with the item-position model was significantly greater than the noise ceiling and compared to the item mixture model ($df = 21, p < 10^{-3}$), and the item mixture model did not reach the noise ceiling. In all displayed plots the lower noise ceilings were significantly greater than zero across participants. Y-axis displays the Spearman's 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: IP: Item-position, II: Item-item, IM: Item mixture.



in the delay or presentation phases of the task in those motor regions.

Since we cannot rule out the possibility that the regions where the item mixture model reached the noise ceiling are not engaged in item rather than sequence representation we excluded those ROIs from further analysis of learning effects (region names displayed in italics in Table 1). It should also be noted that the item-position model and item mixture model are

a priori correlated to each other ($\rho = 0.52$, Fig 3C) since item position effects dominate item mixture components (see *Sequence representation models for RSA* in *Methods*). This means that in regions with significant correlation with the item-position model we should also always see some positive correlation with the item mixture model. However, here we required that the item mixture model should not reach the noise ceiling and be significantly less correlated to data than the item-position model in order to rule out the null hypothesis that the region is representing individual items or finger movements.

The item-position model was the only plausible model for unique sequences

In nine regions in the dorsal visual processing stream the item-position model was both the best-fitting model and 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 [34, 35, 36].

However, there were significant differences in item-position model fits for different phases of the task. For the presentation phase (9.6s duration, four differently oriented Gabors, Fig 2) the item-position 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 unique 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 item-position model was not significantly different from the item mixture model during the response phase (Fig 4, bottom row) suggesting that individual finger movements dominated the voxel patterns in the motor areas.

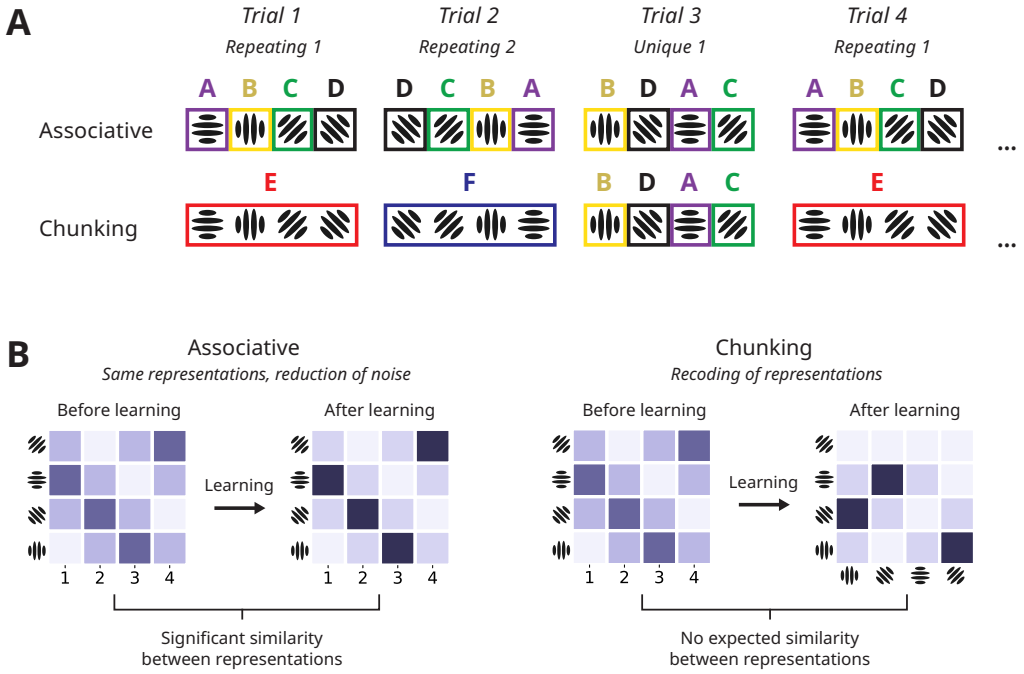
In sum, our analysis of unique sequence representation shows that combining fMRI data with representational similarity models allows more precise quantification of the relationship between cortical activity patterns and models of sequence representation. Using an explicit control model we were able to delineate neural sequence representations from correlated representations of individual items. For unique visual sequences only the item-position 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 [29, 31, 30, 37] and human cortex [38, 39, 40]. Several regions in the dorsal stream also encoded individual items or finger movements. Those regions were excluded from the analysis of learning effects.

Comparison of learning models

The comparison of learning models was carried out in the 9 brain regions where we found significant and sufficient evidence for item-position sequence representation across task phases (Table 1). We tested the data from these regions for contrasting hypotheses made by the associative and chunk learning models.

Fig 5: Learning mechanisms predict different similarity between neural representations. (A) Associative (top) and chunking (bottom) models encode items in individual sequences differently. Differently coloured letters and boxes refer to individual items (top) and chunks (bottom). (B) Measuring similarity between learned and novel sequences. With associative learning (left) learned sequences share the same item codes with novel (unique) ones. Furthermore, learning reduces noise in learned sequence representations. Chunk learning (right) changes item representations so that unique and learned stimuli do not share item codes.



Associative learning

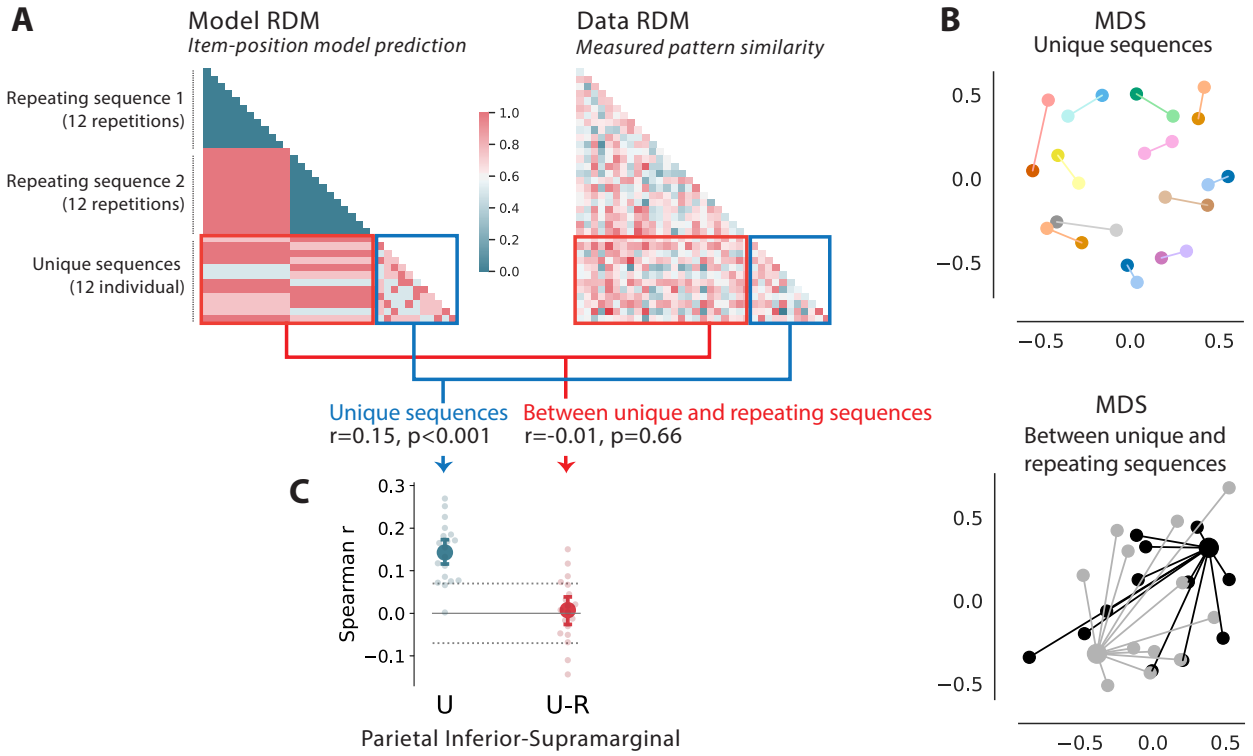
Associative learning predicts that repetition of sequences should leave the individual item representations unchanged but strengthen the weights of individual associations (Fig 5). Formally, the optimal way of encoding how many times particular discrete associations have occurred is given by the Dirichlet-Multinomial model. The full description of the associative sequence learning model and its parameters can be found in *Associative learning* in *Methods*. Intuitively, associative learning can be thought of as changing the weights of associations so that they reflect the frequency of past occurrences. Importantly, since associative learning does not change the representations of individual items themselves, it follows that both unique and repeated sequences should be encoded with the same sequence representation model. We tested this hypothesis by measuring the representational similarity between the unique and repeating sequences (see *Associative learning predictions for RSA* in *Methods* for details).

No evidence for associative learning of items and positions

We found no evidence for the associative learning prediction in any of the ROIs that encoded unique sequences in terms of item-position associations. Specifically, there was no significant correlation between the predictions made by the item-position model and the voxel pattern similarity between repeating and unique sequences (Fig 6A illustrates graphically how this was measured). In all of the 9 regions where the item-position model predicted similarity between unique sequences it failed to predict the similarity between unique and repeating sequences (Fig 6C). This shows that contrary to the predictions of the associative models repeating and unique sequences did not share a common representational code.

The associative learning model also predicts that the learned repeated sequences should be represented with less noise: the repetition of sequences should strengthen the weights of individual associations (see *Associative learning* in *Methods*). Therefore the noise in activity patterns generated by unique sequences should be greater than for repeating sequences and hence the expected similarity *between* repeating and unique sequences should be greater than *within* unique sequences. In other words, it should be easier to find evidence for associative

Fig 6: Unique vs repeating sequence representations. (A) Predictions of the item-position model for the unique sequences (blue squares connected with blue lines) and for the similarity between unique and repeating sequences (red rectangles connected with red lines). The blue squares contain the similarities between all unique sequences. The red rectangles contain the similarities between unique sequences and the two repeating 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 item-position model predictions and data from the parietal inferior-supramarginal gyrus. Top figure shows item-position model predictions for individual unique sequences with lines joining predictions to observed data (the shorter the line the closer data is to the prediction). Bottom figure shows the item-position model's predictions for the distances between unique and two repeating sequences. The predictions for the two repeating sequences are depicted with large dots (black and grey) whilst the data are marked with smaller dots. Both plots use exactly the same MDS projection and display how well the data fit predictions. (C) Model predictions and fit with data in the parietal inferior-supramarginal gyrus: while the item-position model significantly predicts the similarity between unique sequences (blue) it performs at chance level when predicting the similarity between unique and repeating 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.



codes between repeating and unique sequences than for unique sequences alone. Hence the lack of evidence we observe for associative learning cannot be attributed to the lack of fMRI measurement sensitivity. To confirm our assumptions about the effects of noise in learned sequence representations we estimated the fMRI measurement noise for the participants in our task and tested to what degree the noise should be reduced (or signal-to-noise ratio increased) in

the fMRI patterns for the changes to be detectable with the representational similarity analysis (see *Simulation of expected changes in pattern similarity* in *Supplementary information*).

We can further illustrate the difference between the learned and unique sequence representations by visualising the fit between model predictions and data by projecting the representational similarity onto 2D by using multidimensional scaling (MDS). Fig 6B shows how well the item-position model predicts similarity between individual unique sequences (top panel) and between unique and repeating sequences (bottom panel). If, as associative learning predicts, repeating sequences were also encoded positionally but with less noise, then the distances between model predictions and data should be shorter for the unique-repeating comparison (bottom) than for unique sequences only. Fig 6B shows how the opposite is true for the voxel patterns in the parietal inferior-supramarginal gyrus (same scale on both plots).

Predictions of the associative learning model can also be tested behaviourally: increased probability for associations present in the repeating sequences should affect unique sequences where such associations are also present. For example, repeated exposure to a sequence $ABCD$ should also boost $BDCA$ since C appears at the 3rd position in both. We tested this prediction by comparing response times for individual item-position associations in unique sequences: there was no advantage for those associations which were shared with the two repeating sequences ($t = 0.28, p = 0.78, df = 21$).

Chunk learning

If two individual items in a sequence occur next to each more frequently than apart then an optimal learner should infer a model of the environment where those two adjacent items have been generated by a single latent variable. In Bayesian chunk learning participants' internal representations of sequences are inferred as latent variables given the observed sequences:

$$p(\theta|\mathbf{S}) = \frac{p(\mathbf{S}|\theta)p(\theta)}{p(\mathbf{S})}, \tag{1}$$

where θ is the internal latent model of a set of sequences $\mathbf{S} = \{\mathbf{y}_1, ..., \mathbf{y}_m\}$. A chunking model θ_i is defined by two parameters and their probability distributions $p(\mathbf{x}, \mathbf{z}|\theta_i)$, where $\mathbf{x}$ is a set of individual chunks, and $\mathbf{z}$ a set of mappings defining how chunks are arranged together to encode observed sequences. For example, regularities within a set of two sequences $\mathbf{S} = \{ABCD, CDAB\}$ can be expressed by two chunks $\mathbf{x} = \{AB, CD\}$ and their mappings to the observed data $\mathbf{z} = \{((A, B), (C, D)), ((C, D), (A, B))\}$. Here we represent chunks formally as *n-grams* that can take any length up the maximum possible length of a sequence. For example, a four-item sequence $ABCD$ can be represented by a tri-gram ABC and a uni-gram D; or two uni-grams A and B and a bi-gram CD, etc.

Here we used Bayesian model comparison (BMC) [41] to define all possible chunk models and find the optimal model for our task. BMC finds an optimal trade-off between the models' complexity and fit to the data: the optimal model is the one complex enough to generate observed sequences sufficiently well but which does not over-fit the data, which would prevent generalisation to new observations. This approach allowed us to consider all plausible chunking models without pre-specifying their prior probability, as BMC automatically penalises models for their complexity. The full formal description of chunking models, their parameters, and the process of inferring the optimal model can be found in *Chunk learning* in *Methods*.

We estimated the optimal chunking model for every participant at every trial as new sequences were observed. We designed the presentation of repeating and unique sequences so that the optimal model would remain the same for every trial across the experiment: on a trial when a repeating sequence was presented it was encoded with a single four-gram chunk, and when a unique sequence was presented it was encoded with four uni-grams (see *Optimal chunking model* in *Methods* and Fig 9). Hence chunk learning predicts that individual items in the *repeating sequences* would be merged into chunks and therefore items in repeating sequences should be encoded differently from unique ones. It follows that the similarity *between* the unique and repeated sequences should not follow a common representational model (see *Expected representational similarity between encoding models* in *Methods* for details). However, we should be able to dissociate between the neural representations of individual repeating sequences which have

been chunked.

Chunk learning predicts the representation of repeating sequences

As predicted by the chunking model we observed that the representations of unique and repeating sequences were orthogonal to each other: in all of the ROIs the similarity of repeated sequences to unique ones was at chance level. This prediction can be further understood by considering what exactly happens during recoding: any efficient recoding mechanism (i.e. resulting in a shorter code or sparser representation) also changes the dimensionality of the representations [42, 43]. The expected similarity between two representations without shared dimensions should be at chance (see *Expected representational similarity between encoding models* in *Methods* for details). However, we can predict the neural representation of the two repeating sequences from the recoded chunks.

By using voxel patterns corresponding only to the two repeating sequences (*repeating sequence A* vs. *repeating sequence B*) we used a pattern classification approach to test whether brain regions with significant evidence for unique sequence representation (Table 1) also encoded the two repeating sequences as predicted by the optimal chunking model (see *Linear classification of two repeating sequences* in *Methods*). To ensure that the representation of unique and repeating 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 item-position model were included in the classification analysis. This was achieved by running the unique 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 for the two repeating sequences on a subset of voxels which showed significant evidence for the item-position model.

Of the regions which encoded unique sequences (Table 1) we found significant evidence for representations of the two repeating sequences in three brain regions: the parietal inferior-supramarginal gyrus, the postcentral sulcus, and the occipital superior transversal sulcus (Table 2, Fig 7). The classification results were only significant for the presentation phase of the

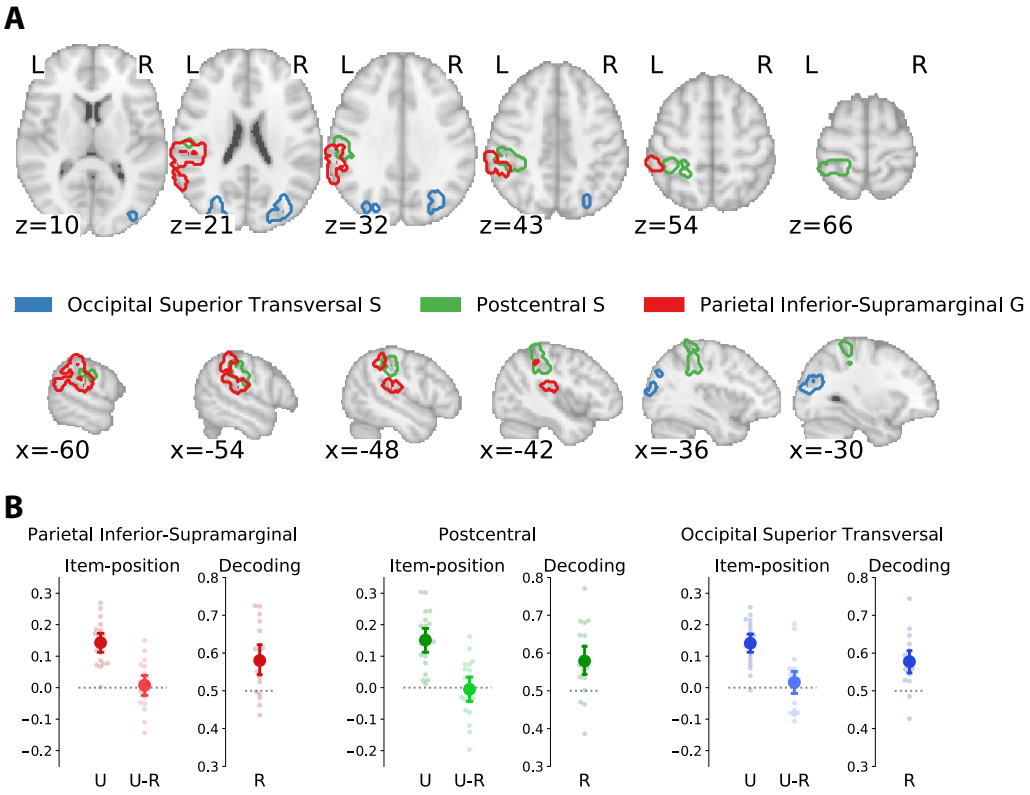
task; we could not decode the two repeating sequences during either the delay or the response phases. This suggests that the decoding results do not reflect individual finger movements corresponding to the two sequences. Similarly, if the decoding analysis reflected a combination of voxel patterns for constituent individual items instead of sequence representations, then we should have seen a significant fit with the item mixture model and unique sequences for these regions (Table 1, Fig 4). Instead, the item mixture model did not reach the noise ceiling in any of these regions.

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

Lobe	Region	Mean accuracy	p
Parietal	Parietal Inferior-Supramarginal G	0.580	0.00069
Parietal	Postcentral S	0.579	0.00074
Occipital	Occipital Superior Transversal S	0.577	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 unique and repeating sequences as predicted by the optimal chunk model (Fig 7B). The representation of unique sequences in all three follows an encoding model where single items are associated with positions in a sequence. However, the same set of voxels also encodes repeating sequences but without single item-position mappings, indicating a change in the representational code, predicted by chunk learning and not by associative learning.

Fig 7: Anatomical localisation of learning effects. (A) Regions which encode both unique and repeating sequences as predicted by the chunk learning model. 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 unique and repeating sequences: plotted for every ROI are the average RSA correlation scores for unique sequences (U) and between unique and repeating sequences (UR), and the average decoding accuracy between two individual repeating sequences (R). Dots: individual subjects' values; error bars around the mean represent bootstrapped 95% confidence intervals.



Discussion

In the current study we contrasted two classes of sequence learning models. First, we considered associative learning that proceeds by changing the signal-to-noise ratio of existing representations. 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: by either reducing uncertainty in the internal representations (associative learning) or reducing the necessary number of associations (chunking). 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 unique and repeating sequences with those predicted by sequence representation models.

Learning induces recoding of sequence representations

We found that *unique* 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 [44, 39]. However, we found no evidence that repeated sequences were also represented positionally, as would be predicted by the associative learning model. Instead, we observed that learning proceeds by recoding the initial stimuli using a different set of codes. Specifically, the similarity between repeated and unique sequences followed predictions of the optimal chunking model in three cortical regions in the parietal lobe: the two repeating sequences shared no representational similarity to the unique ones, while being significantly distinct from each other.

Such flexible recoding of stimuli in response to the changing statistics of the environment is a common and often necessary feature of probabilistic learning models (see Fiser et al. [4] for a review). However, most neural learning models assume that different populations represent different stages of learning: for example, a traditional hippocampal-cortical learning account assumes that the fast acquisition of initial associations is supported by the dynamics of the hippocampus proper while the cortical areas encode the consolidated representations [45]. Here we show that the same cortical region supports both initial and recoded representations.

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 [46] while learned verbal sequences show little indication of positional coding [24]. Furthermore, sequence learning is still possible when the position of items (but not their order) changes from trial to trial [47].

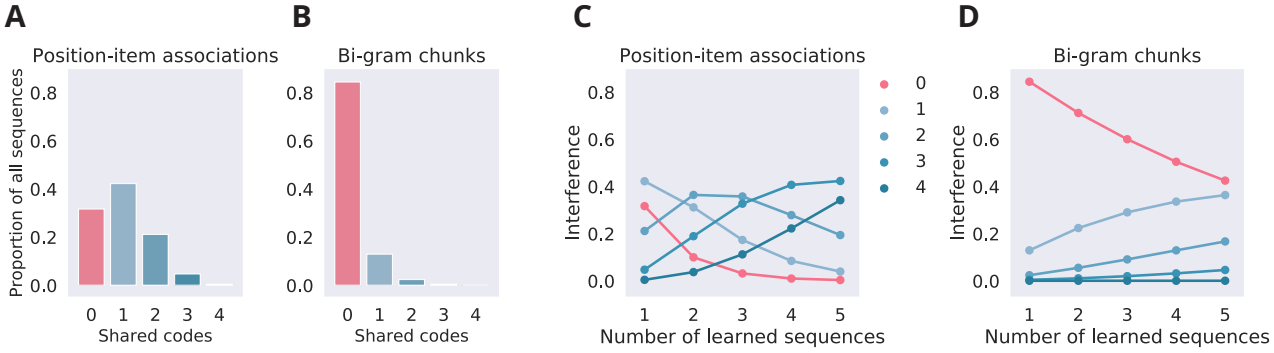
Associative sequence learning is theoretically suboptimal for long-term storage

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 this would result in eight association weights of equal strength to represent four sequences ($ABCD$, $BADC$ and $ABDC$, $BACD$; see *Associative learning of overlapping sequences* in *Supplementary information* for more details). Such a learning mechanism would suffer from catastrophic interference with multiple short sequences of overlapping items. However, most naturally occurring sequences (words or sentences in a language, events in complex cognitive tasks live driving a car or preparing a dish) do not consist of items or events which only uniquely occur in that sequence. Hence an efficient sequence learning mechanism has to be able to learn multiple sequences which are re-orderings of the same items.

The interference resulting from strengthening of individual associations can be quantified for each learning model in terms of shared representations causing such interference. In the item-position model any individual sequence will have an expected similarity to all other possible sequences (Fig 8A). For example, $ABCD$ shares two item-position associations with $DBCA$, and so forth. On the average, any individual 4-item sequence encoded with the item-position model shares two items with 21% of all the other possible sequences, while only 31% share no associations and ca 5% share 3 out of 4 associations (Fig 8A). As more sequences are learned the interference between stored representations will inevitably increase since the number of possible associative codes in the item-position model is fixed (items $\times$ positions). Fig 8C shows that the effective capacity of learning different overlapping sequences with the associative item-position model is approximately five: at that point there are no sequences left which have been

unaffected by learned sequences. Contrastingly, for a chunking model that only uses bi-grams, the expected similarity distribution is markedly different (Fig 8B) resulting in significantly reduced interference between learned sequences (Fig 8D). Note that the bi-gram chunking model is the most limited chunking model possible: any flexible chunking model – such as the one estimated for our participants – will perform significantly better. In sum, associative learning employs very few codes but therefore necessarily loses in the representational power or 'coverage' such learning can provide. Chunking allows emergence of more dissimilar codes which can be used to cover the space of all possible sequences with very little interference.

Fig 8: Interference between sequence representations. (A) Expected number of shared codes (item-position associations, x-axis) in an item-position model for a single 4-item sequences with all other possible 4-item sequences ($n = 256$, allowing repeats) measured as a proportion of sequences sharing the same number of codes (y-axis). (B) Expected number of shared codes for a bi-gram chunk representation. (C) Interference between sequence representations in the item-position model. X-axis displays how many sequences have been learned and lines on the plot display the proportion of other sequences affected by learning as a function of codes shared (panel A). The line numbers (legend) indicate the extent of interference as a number of shared codes. The red line shows the proportion of sequences which have been unaffected by learning. (D) Interference between sequence representations in the bi-gram chunk model.



Multiple and parallel systems for sequence learning

It is important to note that our experimental task is significantly different from standard motor-sequence learning paradigms where learning proceeds through repetition of movements and consolidation can take several hours or days [48, 23, 49]. Here we used a serial recall task where individual sequences are typically learned in as few as 2-4 repeated presentations [50, 51] and learning proceeds even when no recall is attempted [52, 53]. Therefore learning

mechanisms observed in our study are probably more reflective of rapid learning of visual or auditory sequences rather than the slower acquisition of motor skills.

The fast recoding observed in our study is consistent with a model of the hippocampal-cortical learning system where the initial sequence representations are associative, episodic, and created by single exposure [54, 55, 56]. We observe that the initial cortical representations of visual sequences are associative and involve binding individual items to their positions in a sequence. It is thus an open question whether the cortical recoding is mediated by the hippocampal formation: e.g. the initial positional representations of novel sequences are set up in the hippocampus and cortex simultaneously but the interaction between the two set of codes could be modulated by repeated presentations. Alternatively, hippocampal codes could act as pointers to the cortical representations and thus implement chunking [55]. However, it is also possible that the development of chunk-based representations might not need hippocampal involvement at all. The adaptive value of constructing chunk representations in different domains and different modalities might suggest that chunk-based learning should be observed in many brain regions independently.

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).

Fast sequence learning through recoding is very likely only one of the multiple learning processes. Accumulating evidence points to subcortical learning – facilitated by the hippocampal formation and basal ganglia – operating in parallel [57, 58, 59, 60] and the effects of both types of learning can be delineated for a single task in rodents [61]. Therefore we would expect the extent and the exact nature of learning-induced recoding to be dependent on the exact task and its properties.

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 a sequence of items (oriented Gabor patches) displayed in the centre of the screen (Fig 2A). 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 that they should 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 12 times as they learned the mapping of items to button-box buttons. Participants were not informed that there were different types of trials.

Stimuli

All presented sequences were permutations of the same four items (see *Sequence generation and similarity* below on how individual sequences differed from each other). 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).

Structure of the trials

To test our hypotheses we split the 14 individual sequences in to two classes: two of these were repeatedly presented through the experiment (*repeating* sequences, 2/3 of the trials) while the remaining 12 were previously unseen and were only presented once (*unique* sequences, 1/3 of the trials). The two individual repeating sequences were chosen randomly for each participant.

The two *repeating* sequences were also used for training before the scanning experiment (each presented 12 times). This was done to ensure that the two repeating sequences would be 12 times more likely already at the start of the experiment and stay so throughout scanning (see *Optimal chunking model* for details).

To keep the relative probability of repeating and unique sequences fixed throughout the experiment we pseudo-randomised the order of trials so that on the average there was a single

unique sequence and two repeating sequences in three consecutive trials (Fig 2B). This ensured that after every three trials the participant exposure to repeated and unique sequences was the same (2/3 repeated, 1/3 unique sequences).

For MRI scanning we repeated this experimental block twice for every participant so that in a 36-trial scanning session participants recalled each unique sequence once and repeating sequences 12 times each (Fig 2B). Over two scanning sessions this resulted in 48 trials with repeating sequences and 24 trials with unique sequences.

Sequence generation and similarity

We permuted the same four items (oriented Gabor patches) into different individual sequences to resemble sequences in the natural world, which are mostly permutations of a small set of items or events based on the desired outcome (e.g. driving the car, parsing a sentence, etc).

We chose the 14 individual four-item sequences used in the experiment (2 repeating, 12 unique) to be as dissimilar to each other as possible in order to avoid any statistical regularities between individual sequences themselves and instead be able to introduce regularities only through repeating the individual sequences (see *Chunk learning* for details).

We constrained the possible set of individual sequences with two criteria:

1. *Dissimilarity between all individual sequences*: all sequences needed to be at least three edits apart in the Hamming distance space (see *Sequence representation models for RSA* for details on the Hamming distance between two sequences). For example, given a repeating sequence $\{A, B, C, D\}$ we wanted to avoid a unique sequence $\{A, B, D, C\}$ as these share the two first items and hence the latter would only be a partially *unique* sequence. This would allow in chunk learning to encode both sequences with a common multi-item chunk AB.
2. *N-gram dissimilarity between two repeating sequences*: the two repeating sequences shared no items at common positions and no common n-grams, where $n > 1$ (see *Chunk learning* for n-gram definition and details). This ensured that the representations of repeating sequences would not interfere with each other and hence both could be learned to similar

level of familiarity. Secondly, this ensured that for the chunking model the repetitions of these two sequences were optimally encoded with two four-grams since they shared no common bi-grams of tri-grams.

Given these 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 (see *Sequence representation models for RSA*). 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 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 item-position model and item-item model were negative (Fig 3C).

Models of sequence learning

Sequence representation

Sequences are associative codes: they are formed either by associating successive items to each other (item-item associations) or by associating items to some external signal specifying the temporal context for sequences (item-position associations).

In the case of item-item associations the weights of the associations are usually expressed in terms of transitional probabilities [1] forming a 'chain' of associations [62] (Fig 2C). Alternatively, sequences can be formed by associating items to some external signal specifying the temporal context for sequences. This context signal can be a gradually changing temporal signal [63, 64, 65], a discrete value specifying the item's position in a sequence [66], or a combination of multiple context signals [67, 68]. Again, common to all of these models is the underlying association of item representations to the positional signal, forming item-position associations (Fig 2C).

For our sequence processing task (Fig 2) we model the participants' internal sequence representations μ given the presented sequence y as Bayesian inference (Eq 2), where the posterior distribution $p(\mu|y)$ represents a participant's estimate of the presented stimulus, and their response can be thought of as a sample from the posterior distribution:

$$\overbrace{p(\mu|y)}^{posterior} \propto \overbrace{p(y|\mu)}^{likelihood} \cdot \overbrace{p(\mu)}^{prior} \tag{2}$$

Associations between discrete variables – such as items, or items and positions – can be formalised as a multinomial joint probability distribution. The multinomial representation can in turn be visualised as a matrix where each cell describes the probability of a particular item at a particular position (Fig 1).

Formally, every item x in the sequence z is represented by a multinomial variable which can take K states parametrised by a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ which denotes the probability of item

x occurring at any of k positions:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \tag{3}$$

and the whole N-item sequence $\mathbf{z} = (x_1, \ldots, x_N)^T$ is given by:

$$p(\mathbf{z}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}, \tag{4}$$

where the $\boldsymbol{\mu}$ represents the probability of particular item-position associations and hence must satisfy $0 \leq \mu_k \leq 1$ and $\sum_k \mu_k = 1$. Exactly the same formalism applies to item-item associations: we simply replace the set of K position variables with another identical set of N items.

Associative learning

An ideal learner should infer internal representations that reflect the statistics of the environment. In other words, the uncertainty in the inferred representations should match the uncertainty present in the environment. In a learner constrained to operate by associative learning this is achieved by updating the association weights to keep track of how many times items occur at each position.

The optimal way of encoding how many times particular discrete associations have occurred is given by the Dirichlet-Multinomial model. In short, past occurrences of items at certain positions are transformed into probabilities, which reflect the frequency of associations. Hence associative learning can be thought of as changing the weights of associations – $\boldsymbol{\mu}$ parameter in the multinomial model above – so that they reflect the statistics of the environment. This is achieved by deriving $p(\boldsymbol{\mu})$ from the Dirichlet distribution:

$$\boldsymbol{\mu} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{5}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)^T$ denotes the effective number of observations for individual associations. The optimal internal representation of associations for a sequence $\boldsymbol{y}$ is therefore given

by:

$$p(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\alpha}) \propto p(\mathbf{y}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}). \tag{6}$$

We could also use $\boldsymbol{\mu}$ to introduce additional biases into the model (e.g. recency or primacy effects) but since our task has very short sequences and clearly distinctive individual items such additional biases are not significant (see *Behaviour* in *Results*).

In formal terms this means specifying the conjugate prior for the parameter μ of the multinomial prior distribution (Eq 4), which is given by the Dirichlet distribution:

$$p(\mu|\alpha) = \phi \prod_{k=1}^{K} \mu_k^{a_k - 1}, \tag{7}$$

where $0 \leq \mu_k \leq 1$ and $\sum \mu_k = 1$ and ϕ is the normalisation constant. The parameters α_k of the prior distribution can be interpreted as an effective number of observations $x_k = 1$, or in other words, the counts of the value of the sequence position x previously. Effectively, the conjugate prior tracks the item-position occurrence history. Since this model reflects the expected value of item-position associations it is also an optimal model of sequence representation assuming that associations are independent of each other.

Using position-item associations as defined above to encode a set of individual sequences $\mathbf{S} = (BACD, CABD, ABCD)$ will result in a following value for $\boldsymbol{\mu}$ reflecting the position-item counts:

$$\boldsymbol{\mu} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Here matrix rows and columns reflect the items and position variables respectively. Changing or adding new sequences to the set will only change the probabilities or association weights but not change individual items bound by associations. This matrix is visualised for both item-position and item-item associations in Fig 3B.

Note that the assumption that sequential associations are independent of each other does not seem to hold for human memory of longer sequences. For example, for long sequences which exceed the short-term memory span, first items are generally more accurately recalled compared to the rest reflecting a primacy bias [69]. Similar bias has been observed for the last items in supra-span sequences (recency effect). However, since such additional biases would be constant across sequences – i.e. advantage for the first item is always the same all other things equal – their effects are orthogonal to the frequency of associations and hence not modelled here. Furthermore, we chose the stimuli in our task – short, four-item sequences – so that they would easily fit within participants short term memory and therefore not elicit such additional biases.

Chunk learning

Bayesian model comparison

We want to estimate the posterior probability distribution of chunking models θ given the observed data D:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \tag{8}$$

and choose the model with the highest posterior probability:

$$\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}[p(\theta|D)].$$

Since Bayesian model comparison (BMC) implements an inherent Occam's razor which penalises models in terms of their complexity (see below) we assign all models equal prior probabilities $p(\theta)$. Therefore the posterior probability of any model is proportional to model *evidence*:

$$p(D|\theta_i) = \int p(\mathbf{S}|\mathbf{w}, \theta_i)p(\mathbf{w}|\theta_i)d\mathbf{w}, \tag{9}$$

where $\mathbf{S}$ is a set of sequences (data), θ_i a particular chunking model and $\mathbf{w}$ its parameter values. Intuitively, to estimate *evidence* for any model we need to evaluate its complexity as defined

by its parameters $\mathbf{w}$ and their probability distributions $p(\mathbf{w}|\theta_i)$, and how well the model fits the data $p(\mathbf{S}|\mathbf{w}, \theta_i)$. By combining the model complexity and data fit we can rank all possible models in terms of their respective evidence $p(D|\theta_i)$. The model with the greatest evidence is also the model with maximum *a posteriori* probability since we assume equal prior probabilities across models.

Chunking model

A chunking model θ_i is defined by two parameters and their probability distributions $p(\mathbf{x}, \mathbf{z}|\theta_i)$, where $\mathbf{x}$ is a set of individual chunks and $\mathbf{z}$ a set of mappings defining how chunks are arranged together to encode observed sequences.

Set of chunks

We represent chunks formally as n-grams (used from hereon synonymously with the term 'chunk') that can take any length up the maximum possible length of a sequence to be encoded. For illustrative purposes we denote the individual items in our sequences here with letters: a four-item sequence of Gabors can be written as $ABCD$ and in turn be represented by a tri-gram ABC and a uni-gram D. For 4-item sequences the set of all possible n-grams has the number of P members as the sum of partial permutations of $n = 4$ items taken $k = \{1, 2, 3, 4\}$ at a time:

$$P_N = \sum_{k=1}^{n} \frac{n!}{(n-k)!} = 64$$

A set of chunks $\mathbf{x}$ comprises J n-grams where each constituent n-gram c appears only once: $\mathbf{x} = \{c_1, ..., c_J\}$, and $1 < J < P_N$; for example $\mathbf{x} = \{AB, BA, A, B, CDA, ACDB\}$. Each individual n-gram has a probability inversely proportional to its combinatorial complexity. Specifically, the probability of a particular n-gram c_j is proportional to the reciprocal of the number of partial permutations of $n = 4$ items taken k at a time:

$$p(c_j) = \alpha \frac{1}{\frac{n!}{(n-k)!}}, \tag{10}$$

where k is the length of the n-gram and α is the normalising constant. For example, there are 4 possible uni-grams for a 4-item sequence, but 12 bi-grams, 24 tri-grams, etc. Hence the probability of a uni-gram is 3 times greater than a bi-gram and so forth. This also captures the intuition that longer and more complex chunks should be less likely than simple chunks. We also assume that chunks are independent each other and hence the probability of a set of n-grams $\mathbf{x}$ defined by the chunking model is the product of its constituent chunk probabilities:

$$p(\mathbf{x}|\theta_i) = \prod_{j=1}^{J} p(c_j). \tag{11}$$

Mappings between chunks

The second parameter of the chunking model describes how individual n-grams are combined together to encode the observed sequences. For example, given a single sequence $ABCD$ and a set of n-grams $\mathbf{x} = \{AB, BC, CD, A, B, C, D\}$ we can encode the data as (AB, CD) or (A, BC, D) as both mappings are capable of representing the observed data without error.

For any 4-item sequence there are eight possible ways n-grams can be linked together to reproduce the observed sequence. These mappings can be described as a set of eight tuples $\mathbf{Z} = \{\mathbf{g}_1, \ldots, \mathbf{g}_8\}$, where each tuple defines $F \leq 4$ links $\mathbf{g}_i = ((l_1), \ldots, (l_F))$ that exhaustively

define all possible n-gram parses of a 4-item sequence:

$$\begin{aligned}\mathbf{Z} =\{&((1),(2,3),(4)),\\ &((1),(2,3,4)),\\ &((1,2,3),(4)),\\ &((1,2),(3),(4)),\\ &((1),(2),(3),(4)),\\ &((1),(2),(3,4)),\\ &((1,2,3,4)),\\ &((1,2),(3,4))\}.\end{aligned}$$

For example, given a sequence $ABCD$, the first tuple in the set $\mathbf{g}_1 = ((1,),(2,3),(4,))$ corresponds to linking three individual n-grams as $((A),(B,C),(D))$. The mappings $\mathbf{g}_i$ in $\mathbf{Z}$ differ in terms of how many links are required to encode the sequence: for example, the first mapping comprises three links, the second two, and the fifth four. The probability of each mapping is a product of it's individual link probabilities:

$$p(\mathbf{g}_i) = \prod_{j=1}^{F} p(l_j) = \eta^F, \tag{12}$$

where F is the number of links in the mapping $\mathbf{g}_i$ and η is a probability of each link which we assume to be constant (the reciprocal of the number of possible links). Such inverse relationship between the probability of a mapping and its length captures the intuition that complex mappings between multiple n-grams should be less likely than simple ones. Note that for longer sequences a different relationship might be defined as the ability to combine chunks is limited by human short term memory capacity which sets an effective limit to the length of sequence that can be encoded [70, 71, 72].

For a particular model θ_i the mapping parameter $\mathbf{z}$ defines how n-grams are combined

together to generate observed sequences. For example, consider two models and a set of sequences $\mathbf{S} = \{ABCD, ABCD, CDAB\}$. Both models use the same set of n-grams $\mathbf{x} = \{AB, CD, A, B, C, D\}$, but encode the observed sequences differently:

$$\mathbf{z}_1 = \{((A,B),(C,D)),((A,B),(C,D)),((C,D),(A,B))\}$$
$$\mathbf{z}_2 = \{((A),(B),(C),(D)),((A),(B),(C),(D)),((C),(D),(A),(B))\}$$

Although these two models are equally likely in terms of the chunks they employ, their mappings have different probabilities. The probability of mappings defined by a particular model is equal to the product of mappings for individual sequences:

$$p(\mathbf{z}|\theta_i) = \prod_{i=1}^{M} p(\mathbf{g}_i), \tag{13}$$

where M is the number of mappings (sequences encoded).

For any model θ_i the probability of both parameters – n-grams and mappings – are assumed to be independent of each other and therefore the probability of any particular model is the product their parameter probabilities:

$$p(\mathbf{x}, \mathbf{z}|\theta_i) = \prod_{j=1}^{J} p(c_j) \prod_{i=1}^{M} p(\mathbf{g}_i), \tag{14}$$

where J and M are the number of n-grams and mappings respectively. Therefore every model and its two parameters propose an encoding based on some chunks (e.g. examples above), which can subsequently compared to the observed data.

Model fit

In Bayesian inference the model fit is defined by the likelihood function which evaluates how likely is that the observed data was generated by a particular model – in our case:

$$p(\mathbf{S}|\theta_i) = p(\mathbf{S}|\mathbf{z}, \mathbf{x}, \theta_i),$$

where $\mathbf{x}$ is a set of n-grams and $\mathbf{z}$ is a set of discrete mappings which define how individual n-grams are combined to encode the observed data $\mathbf{S}$. Intuitively, the likelihood of a model θ_i quantifies how easy or difficult it is to generate all observed sequences using a set of n-grams and mappings as specified by the model.

Commonly, the likelihood of a model is measured in terms of the distance between model predictions and the observed data: for example, we could use a between-sequence distance metric (such as the Levenshtein or Hamming distance) to compute the distances between the observed sequences $\mathbf{S}$ and the set of sequences defined by a particular model's parameters (n-grams and mappings). However, here we only consider models that are capable of encoding the observed data: e.g. for a set of two sequences $\mathbf{S} = \{ABCD, DBAC\}$ we only consider chunks like $\mathbf{x} = \{AB, CD\}$ or $\mathbf{x} = \{A, B, C, D\}$, but not $\mathbf{x} = \{CA, DD\}$; and the same with mappings. There are two reasons for this: first, the space of possible models that correctly encode the observed sequences is already quite large. For example, in our study we use 14 individual sequences. As any individual 4-item sequence can be encoded with 8 different mappings (see *Chunking model* above), it follows that a set of 14 sequences can be encoded with 8^{14} different mappings. Similar combinatorial expansion applies for the number of possible sets of n-grams. Second, the models that cannot even theoretically fit the data are inevitably less likely than models which do. Therefore by constraining ourselves to the subspace of data-matching models we explore the domain of most probable models. This constraint also follows an ecological rationale: chunks are assumed to be inferred from the regularities present in the data, hence there is no reason to consider latent variables that cannot be mapped onto the observed variables.

Model evidence

Model evidence (Eq 9) combines previously described measures: model complexity in terms of the probabilities of its parameters and model fit. Here we only consider models which fit the observed data perfectly: evaluating model evidence is therefore reduced to estimating model complexity for data-fitting models. The model with greatest evidence – and therefore the one with maximum a posterior probability – is the one which encodes the set of observed sequences with the least complex model.

Importantly, the two model parameters – set of n-grams and mappings – make contrasting contributions to model complexity: an optimal model will need to find a trade-off between the number of n-grams it comprises and the complexity of the mappings. For example, a set of four individual uni-grams $\mathbf{x} = \{A, B, C, D\}$ can encode any of the 14 sequences in our task, but all of the mappings need to be maximally complex, each involving four links between the n-grams. Such a model would have a very simple set of chunks but would require complex mappings to encode the observed sequences. In the other extreme, consider a model where each individual sequence is encoded with a single four-gram and therefore would require very simple mappings (each n-gram to each individual sequence, i.e. four times less complex per sequence than the uni-gram model). However, the such a set of 14 four-grams is by definition more complex and therefore less probable than a set of four simple uni-grams.

The two model parameters – set of n-grams and mappings – can therefore be intuitively thought of as the model's *codes* and the *encoding* it specifies. The Bayesian model comparison mechanism guarantees that the model with the greatest evidence – the optimal model – will define an ideal trade-off between the complexity of the codes and the encoding it produces. This trade-off can be visualised by displaying the model evidence as a sum of their negative log probabilities: Fig 9B illustrates the trade-off between the codes and the encoding for several possible chunking models given a set of two repeated sequences.

Optimal chunking model

The optimal model can be estimated by either randomly sampling the parameter distributions or using a systematic approach. Here we found the optimal model by only considering models that result in parsing the set of sequences into chunks and creating a ranking based on model evidence as described above. Furthermore, we designed the experiment so that the only regularities between the sequences that could be encoded with common chunks arose from repeating the same two sequences: this ensured that we effectively knew the optimal model beforehand.

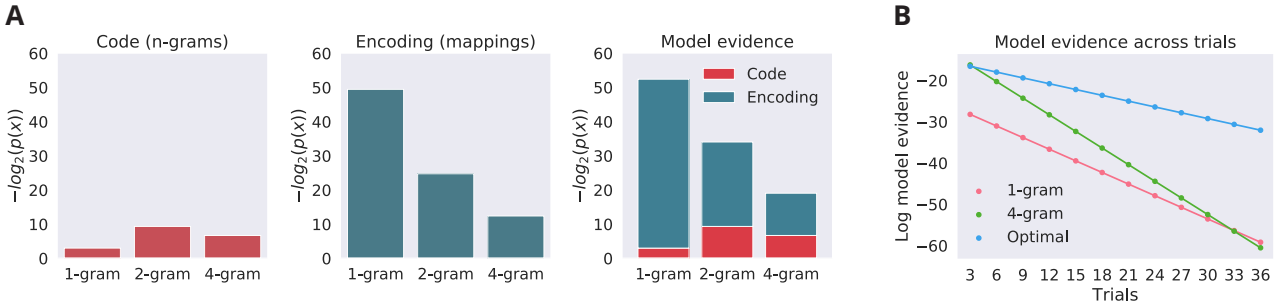
To recall, our task included 14 individual sequences made maximally dissimilar to each other with two of them repeated on 2/3 of the trials. We first presented the two repeating sequences 12 times each during the practice session immediately before the experiment. Since those two sequences shared no common bi-grams or tri-grams (see *Sequence generation and similarity*) the only efficient chunk encoding for the set of repetitions of those two sequences comprised just two four-item chunks (Fig 9A). Using 12 repetitions in the training session was enough to make the four-gram representation the most likely one (Fig 9B). Therefore at the start of the experiment (before the first trial $t = 1$) the optimal chunking model had a set chunks defined as:

$$\mathbf{x}_{t=0}^{MAP} = \{CADB, DCBA\},$$

and a set of mappings $\mathbf{z}_{t=0}^{MAP}$ where the set of 24 sequences (made up of just two individual sequences) were encoded with the same one-on-one mappings.

We proceeded to estimate the optimal chunking model at every trial t as the set of sequences was updated with newly observed stimulus. For this purpose we kept the statistical structure of the sequences fixed across the experiment: otherwise an optimal model on trial one would be different to the one on the last trial. Therefore we organised the order of trials so that on the average there was a single unique sequence and two repeating sequences in three consecutive trials. This ensured that after every three trials the participant exposure to repeated and unique sequences was roughly the same.

Fig 9: Optimal chunking model. (A) Evidence for three models and their components at the beginning of the scanning experiment, when participants had seen the two repeating sequences 12 times each during the practice session. The three models use only single type n-grams: the 1-gram model encodes sequences using four single-item n-grams, 2-gram model with two bi-grams, and the 4-gram model with a single four-gram. The left panel shows the probability of the set of n-grams (code) each model specifies in terms of their negative log values. The centre panel shows the probability of their mappings (encoding) and the right panel the combination of the two into model evidence. The blue and red parts of the model evidence bar represent model code (n-grams) and encoding (mappings) probabilities in terms of their negative logs and the total length of the bar displays the model evidence as their sum. This allows intuitive visualisation of the code-encoding trade-off calculated by the Bayesian model comparison. The 4-gram model is the optimal model at the start of the experiment. (B) Model evidence across trials. X-axis shows the trial number and y-axis shows the log of model evidence. The optimal model is inferred at every trial; the 1-gram model encodes sequences only with four uni-grams, and the 4-gram model only uses four-grams. Note that at the beginning of the experiment the 4-gram model is equivalent to the optimal model: however, as new sequences are presented the optimal model encodes new data with shorter chunks (uni-grams) while the 4-gram model encodes new unique sequences with four-grams. Note that as new data is observed the evidence for any particular model decreases as the set of data becomes larger and the space of possible models increases exponentially. Also note that the log scale transforms the change of evidence over trials into linear form.



Since the unique sequences shared no significant statistical regularities with each other or with the repeating sequences they could have not been efficiently encoded with common n-grams ($n > 1$). Therefore, at trial $t = 1$ the optimal model to encode the previously seen repeating sequences and the new unique one included the previously inferred two four-grams and additionally four single item uni-grams:

$$\mathbf{x}_{t=1}^{MAP} = \{CADB, DCBA, A, B, C, D\}.$$

Since we kept the statistical structure of the sequences fixed across the experiment this ensured that the optimal model would remain the same for every trial across the experiment: on a trial when a repeating sequence was presented it was encoded with a single four-gram

chunk, and when a unique sequence was presented it was encoded with four uni-grams, as shown on Fig 9B.

Representational similarity analysis (RSA)

Sequence representation models for RSA

We defined three models of sequence representation which either encode sequences as associations between items and temporal order positions (*item-position model*, Fig 3A-B, first column), pairs of successive item-item associations (*item-item model*, Fig 3A-B, second column), or as a weighted mixture of item representations (*item mixture model*, Fig 3A-B, third column). The item mixture model controls for the null-hypothesis that fMRI voxel activity could be explained by the superposition of patterns for constituent individual items.

Item-position associations

When sequences are represented as item-position 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:

$$D_H(\mathbf{y}_j, \mathbf{y}_l) = \sum_{i=1}^{k} |x_j^i - x_l^i| \tag{15}$$

$$x_j^i = x_l^i \Rightarrow 0 \tag{16}$$

$$x_j^i \neq x_l^i \Rightarrow 1 \tag{17}$$

where x_j^i and x_l^i are the i-th items from sequences $\mathbf{y}_j$ and $\mathbf{y}_l$ of equal length k respectively. Consider two sequences $ABCD$ and $CBAD$: 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).

We use the between-sequence similarity as defined by the Hamming distance as a prediction

about the similarity between fMRI activation patterns: if sequences are coded as item-position associations then the similarity of their corresponding neural activity patterns, all else being equal, should follow the Hamming distance. This allows us to test whether a particular set of voxels encodes information about sequences using an item-position model. *Representational similarity analysis of fMRI activity patterns* below provides the details of the implementation.

Item-item associations

Here we again use n-grams as associations between multiple consecutive items to define sequences as pairwise item-item associations (Fig 1B): a four-item sequence $ABCD$ can be unambiguously represented by three bi-grams AB, BC, CD so that every bi-gram represents associations between successive items. The bi-gram representation of item-item associations can be used to derive a hypothesis about the similarity between sequences: the between-sequence similarity is proportional to how many common item pairs they share. For example, the sequences FBI and BIN both could be encoded using a bi-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 bi-grams are retained between two sequences:

$$S_C(S_i, S_j) = \text{card}(C_i \cap C_j) \tag{18}$$

where C_i and C_j are the sets of n-grams required to encode sequences S_i and S_j respectively so that $\text{card}(C_i \cap C_j)$ denotes the cardinality of the union of two sets of n-grams (i.e. the number of elements in the resulting set). 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. For bi-grams this gives:

$$C_i = \{i = 1, \ldots, k-1 \ : (x_i, x_{i+1})\}$$

where C_i is a set of all possible adjacent n-grams from sequence S_i of length k so that every

bi-gram is a pair (tuple) of consecutive sequence items (x_i, x_{i+1}). Similarly for a set of n-grams C from any sequence of length k:

$$C = \{i = 1, \ldots, k - (n-1) \ : (x_i, \ldots, x_{i+(n-1)})\},$$

where n is the length of n-gram. Effectively, the n-gram similarity counts the common members between two n-gram sets. Given sequence length k this similarity can accommodate n-grams 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:

$$D_C = 1 - \gamma \ \mathrm{card}(C_i \cap C_j) \tag{19}$$

where γ is a normalising constant:

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

Effectively, the n-gram distance D_C counts the common members between two n-gram sets. We then used the bi-gram distance measure to derive sequence representation predictions for item-item association models.

The prediction made by the n-gram distance D_C is fundamentally different from the prediction made by the Hamming distance D_H (Eq 17): the n-gram distance assumes that sequences are encoded as item-item associations (Fig 1B) whilst the Hamming distance assumes sequences are encoded as item-position associations (Fig 1A). Fig 3B shows the similarity between individual sequences according a bi-gram model.

To understand why the item-position and item-item models make inversely correlated predictions, consider again the example given above: two sequences of same items FBI and BIF are similar from a bi-gram perspective since both could be encoded using a bi-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

(bi-grams) and are hence dissimilar form a bi-gram perspective but have both F at the first position and hence somewhat similar in terms of the item-position model (Hamming distance).

Item mixture

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 $ABCD$ 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:

$$M = \mathbf{I}\beta \tag{20}$$

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 $\beta = \{0, 1/6, 1/3, 1/2\}$ represents a recency-weighted slope over individual sequence items. The reason we only tested for the 'recency mixture' is that the distances between mixtures only depend on the absolute value of the slope of the mixture coefficients over the

items. In other words, an RDM derived with a recency-based item mixture predicts the same similarity between voxel patterns as an RDM derived with a primacy based mixture given the absolute value of the gradient slope remains the same. Here we chose a middle point between two extreme slope values: all the mixtures become the same when the slope is horizontally flat and only a single item contributes when the slope is vertical. See *Item mixture model parameters* in the *Supplementary information* for more details and a worked example.

Distances between two item mixture representations M_i and M_j (Eq 20) of sequences S_i and S_j were calculated as correlation distances:

$$D_I(S_i, S_j) = cdist(M_i, M_j). \tag{21}$$

Representational similarity analysis of fMRI activity patterns

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

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

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

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 unique sequences. $D(N_i, N_j)$ is the distance measure corresponding to one of the three sequence representation models:

1. The item-position model: Hamming distance (Eq 17)
2. The item-item model: bi-gram distance (Eq 19)
3. The item mixture model: the item mixture distance (Eq 21)

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 unique sequences:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \dots & a_{1,M} \\ \vdots & \ddots & \vdots \\ a_{M,1} & \dots & a_{M,M} \end{bmatrix}, \tag{22}$$

$$a_{ij} = cdist(P_i, P_j) = 1 - \frac{(P_i - \mu_{P_i}) \cdot (P_j - \mu_{P_j})}{||(P_i - \mu_{P_i})||_2 ||(P_j - \mu_{P_j})||_2} \tag{23}$$

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 unique sequences N_i and N_j. As shown by Eq 23, 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:

$$r_{r,p} = \rho(\mathbf{rS}_{r,p}, \mathbf{rA}_{r,p}) = \frac{\mathbb{E}[(\mathbf{rS}_{r,p} - \mu_{\mathbf{rS}_{r,p}})(\mathbf{rA}_{r,p} - \mu_{\mathbf{rA}_{r,p}})]}{\sigma_{\mathbf{rS}_{r,p}} \sigma_{\mathbf{rA}_{r,p}}} \tag{24}$$

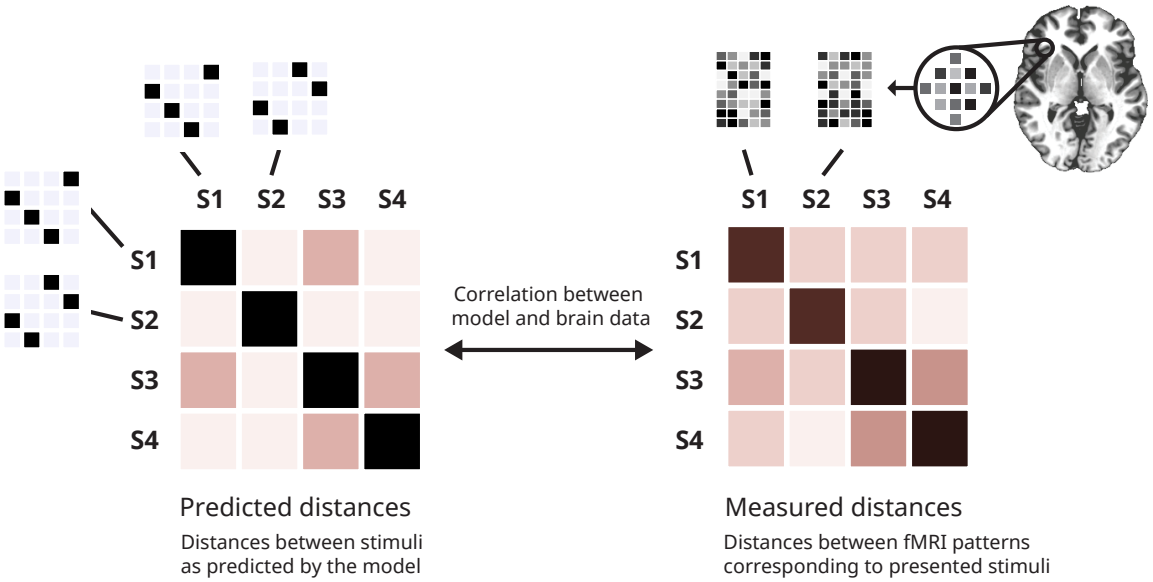
where ρ is the Pearson correlation coefficient applied to the ranks $\mathbf{rS}$ and $\mathbf{rA}$ of data $\mathbf{S}$ and $\mathbf{A}$ respectively.

Finally, to identify which ROIs represented unique 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 10.

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 estimate of the noise ceiling is necessary to indicate how much variance

Fig 10: Representational similarity analysis. Left: model prediction expressed as an RDM of pairwise between-stimulus distances (e.g. Hamming distance for the item-position model). The small matrices refer to the representations of individual sequences in the matrix form (as shown on Fig 1). Right: measured RDM of voxel activity patterns elicited by the stimuli. The small matrices are illustrative representations of voxel × trials matrix from an arbitrary brain region. 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*).



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 22, 23) 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.

We also tested whether a model's noise ceiling was significantly greater than zero. We first Fisher transformed individual Spearman's rank-order correlation values and then performed a one-sided t-test for the mean being greater than zero. The 5% significance threshold for the t-value was corrected for multiple comparisons as described in *Significance testing*. For more information about the noise ceiling please see Nili et al. [73].

Associative learning predictions for RSA

Associative learning makes two predictions: learning doesn't change individual item representations and learning reduces noise in sequence representations. These hypotheses can be tested by measuring the similarity between neural activation patterns elicited by unique and repeating sequences.

Noise in sequence representations can be estimated by assuming that the voxel pattern similarity $\mathbf{A}$ is a function of the 'true' representational similarity between sequences $\mathbf{S}$ plus some measurement and/or cognitive noise ν: $\mathbf{A} = \mathbf{S} + \nu$. Here the noise ν is 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 unique sequences ν_N should be greater than for repeating sequences ν_L. To test this we measured whether the activity patterns of repeating sequences were similar to unique sequences as predicted by the Hamming distance. The analysis followed exactly the same RSA steps as above, except instead of carrying it out within unique sequences we do this between unique and repeating sequences. First, we computed the Hamming distances between individual repeating and unique sequences $\mathbf{S}_{U,R}$, next the corresponding voxel pattern similarities $\mathbf{A}_{U,R}$ and finally computed the Spearman's rank-order correlation between the stimulus and voxel pattern RDMs exactly as above (Eq 24). If this measured correlation is significantly greater than the one within unique sequences ($r_{U,R} > r_U$) across participants, it follows that the noise level in repeating representations is lower than in unique representations. This analysis was carried out for all task phases and in all ROIs where unique 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 $r_{U,R}$ is less than the significance threshold ($p < 10^{-4}$; see *Significance testing* below). This means that repeating sequences are not

represented as item-position associations in this ROI and hence the test for noise levels is meaningless.

2. Significant correlation, but consistently smaller across participants than the within-unique sequences measure: $r_{U,R} < r_U$. repeating sequence representations are noisier than unique sequence representations.

3. Significant correlation, but consistently greater across participants than the within-unique sequences measure: $r_{U,R} > r_U$. repeating sequence representations are less noisy than unique sequence representations.

To confirm that our assumptions regarding the effects of noise on sequence representation were correct we estimated the fMRI measurement noise for the participants in our task and tested to what degree the noise should be reduced (or signal-to-noise ratio increased) in the fMRI patterns for the changes to be detectable with the representational similarity analysis. The details of these simulation can be found in *Simulation of expected changes in pattern similarity* in *Supplementary information*.

Chunk learning predictions for RSA

The chunk learning model predicts that the repetition of individual sequences should recode individual associations for those sequences. Specifically, an optimal chunk model for our task would represent the two repeating sequences with single chunks each, but leave the items in unique sequences the same (see *Optimal chunking model* above). The similarity between unique and repeating sequences should therefore be at chance but the representations for the two repeating sequences should be clearly dissociable from each other.

Specifically, this leads to the following predictions about the similarity of neural sequence representations:

1. Unique and repeating sequences use different item representations: the similarity between unique and repeating sequences as predicted by a sequence representation model should not be significant and the mean correlation value $r_{U,R}$ should be zero. See *Expected*

representational similarity between encoding models below for details why the expected similarity between two different chunking models should be at chance.

2. Each of the two repeating sequences is consistently represented by a different individual chunk which will make it possible to dissociate between the representations of the two individual sequences.

The reason why the similarity predictions don't change for unique sequences even after recoding can be illustrated when considering that even if single item representations change (from 'original' items to 'new' single item chunks) these new items still need to be bound together into a new sequence. Given that both old and new codes correspond to single items these associations should remain the same and hence predicted by a sequence representation model as above (*Representational similarity analysis of fMRI activity patterns*).

Each individual repeating 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 also encoded repeating sequences by using support vector machine (SVM) classification on voxel patterns pertaining to either of the repeating sequences. The details of the SVM testing, training, and significance assessment procedures are described below.

Expected representational similarity between encoding models

Chunking recodes initial sequence representations so that individual items are merged together into new codes. What should be the expected similarity between such new codes and the initial ones?

We assume that initial sequences representations are associations involving individual items, such as the item-position model (Fig 1). The similarity of these initial representations to new recoded chunks then depends on whether both sets of item representations overlap in some way. For example, a code for a single individual item A and a chunk AB could be similar since they both share the same 'original' item A: a neural unit encoding individual items A might hence participate in encoding newly created chunks involving A. However, such a chunk encoder would not carry out efficient recoding.

First, any efficient recoding mechanism should reduce the amount of information (e.g. bits) that is required to store the data without loss. In other words, for an efficient coding scheme the probability of the 'event' is proportional to the amount of information required to encode it (e.g. number of bits). This is only possible if, after recoding, the codes for the frequent events (repeated sequences) are shorter than the initial ones. This means that efficient recoding must change the dimensionality of representations: e.g. from item-position dimensions to chunk-position dimensions. Representations which do not share dimensions should be on the average orthogonal to each other: for example, the dimensionality of linear representations can be estimated by orthogonalising the measurements with respect to their eigenspectrum. See [42, 43] for a detailed review and discussion on this topic.

Lastly, chunks which share the similarity space with initial individual item associations would not be capable of overcoming the combinatorial problem of associative learning, which fails when multiple overlapping sequences have to be learned. This is detailed in *Associative sequence learning is theoretically suboptimal for long-term storage* in *Discussion*.

Data analysis

Behavioural measures

Significant differences in behavioural measures across participants were evaluated with a t-test for dependent measures. We chose not to inverse-transform reaction time data following recent advice by Schramm and Rouder [74] (see also Baayen and Milin [75]).

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 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 [76]. 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 [77, 78], which is based on *Nipype* 1.1.3 [79, 80]. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using `N4BiasFieldCorrection` [81, 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` [82, 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 [83]. Spatial normalisation to the ICBM 152 Nonlinear Asymmetrical template version 2009c [84] was performed through nonlinear registration with `antsRegistration` [85, 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` [86, 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 [87]. 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` [88, FSL 5.0.9]. The BOLD time-series were slice-time corrected using `3dTshift` from AFNI [89] 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. [84]) 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 [90]. 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 2A) 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 et al. [91]). 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 [92] and automatically identified and segmented for each participant using `mri_annotation2label` and `mri_label2vol` (FreeSurfer).

Linear classification of two repeating sequences

We ran a binary classification of sequence identity (*repeating sequence A* vs. *repeating sequence B*) using only data from trials where a repeating sequence was presented: 24 trials of *repeating sequence A* and 24 trials of *repeating 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 regularisation 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 [93] 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 unique 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 to compute 1000 mean RSA correlation coefficients (Eq 24). For the classification analysis we permuted the correspondence between the test labels and data to compute 1000 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 mean scores (with replacement) from each participant's permuted distribution. The number of random samples for the group mean distribution was dependent on the significant probability threshold: we took $n = 10/\alpha$ samples so that the number of samples was always an order of magnitude greater than the required resolution for the group chance 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*.

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, and Jörn Diedrichsen for insightful comments on the earlier version of the manuscript. This research was supported by the Medical Research

Council UK (SUAG/012/RG91365).

Supplementary information

Item mixture model parameters

There are a number of meaningful ways individual items could contribute to the mixture. Although we have chosen a 'recency mixture' where the most recent item contributes the most, we could have also used a 'primacy mixture' with exactly the opposite slope of mixture contributions. The reason we only tested for the 'recency mixture' is that both recency and primacy models predict the same similarity between individual sequences in our task. In other words, a representational dissimilarity matrix (RDM, Eq 22) derived with a recency-based item mixture predicts the same similarity between voxel patterns as an RDM derived with a primacy based mixture (if the absolute value of the gradient slope remains the same). For a detailed explanation see the example below.

We could have chosen any of the infinite slope values across positions. However we chose a middle point between two extreme slope values: all the mixtures become the same when the slope is horizontally flat and only a single item contributes when the slope is vertical. We could have obtained an estimate of the coefficients from analysing the individual finger movement representations since the mapping between items and fingers was randomised across the participants. However, here our focus was on sequence representations and therefore we felt a null hypothesis representing an average in the space of possible mixtures was enough.

Worked example

We assume a recency mixture model where contributions increase with sequence position as $\beta = [0, 1/6, 1/3, 1/2]$, then we can represent a sequence as an item mixture in our task by indicating the proportion of each four items $[A, B, C, D]$ in the mixture, e.g.: $[C, A, D, B]$ as $[A : 1/6, B : 1/2, C : 0, D : 1/3]$, $[B, D, C, A]$ as $[A : 1/3, B : 0, C : 1/6, D : 1/2]$ and so forth. Given that item representations do not change from sequence to sequence and hence all mixtures would be equal if the coefficients β were equal across all items (e.g. uniformly 1/4) the distances between the resulting mixture representations are determined by the vector of coefficients. For

example, the euclidean distance between $[A, B, C, D]$ and $[C, A, D, B]$ as mixtures (given $\beta = [0, 1/6, 1/3, 1/2]$) is the euclidean distance between the two four-dimensional mixture vectors:

$$d = EuclideanDist([1/6, 1/2, 0, 1/3], [1/3, 0, 1/6, 1/2])$$

Assuming the gradient β has always the same number of unique values then the distance between such 4D points depends only on the absolute value of the gradient slope and not the direction of it (positive or negative slope). This should be evident when one considers that in our task all sequences have always exactly the same four items and hence mixture contributions are directly proportional to the ordering of the same four items.

For a simulation how the mixture model similarity prediction does not depend on the direction of the slope (recency vs primacy) see the Jupyter Notebook (*model_mixture*) at our code repository.

Simulation of expected changes in pattern similarity

Our hypothesis on the effects of associative learning assumes noise reduction directly at the level of representational dissimilarity matrix (RDM, Eq 22). Diedrichsen et al. [94] have pointed out that as most distance estimates are based on the product of random variables, the resultant noise variance in the distance estimates gets more complicated than the model we have assume here. To address this issue we investigated to what degree the noise should be reduced (or SNR to be increased) in the learned pattern in order for the changes to be detectable in the RDM. Specifically, we carried out a series of simulations to assess how pattern similarity distances change according to the reduction of noise in the activity patterns for the learned sequences. For example, when the measurement noise is already high, a certain amount of noise-reduction in learned sequences would not be visible in the estimated distance measures.

We simulated the predicted effects of associative learning which assume that (1) neural sequence representations remain the same with learning but (2) their SNR changes proportional to the SNR change observed in the behavioural responses. Therefore such change should also be detected in the fMRI data.

Briefly, we first transformed the behavioural change accompanying sequence learning (sig-

nificant reduction in manual response times, see *Behaviour* in the *Results* section of the main manuscript) into the change in the internal representations of sequences. Formally, we assumed that manual response times are proportional to noise in the representation distribution: as noise increases so do the response times. This leads to two representational noise estimates for both unique and repeated sequences which were then transformed into expected voxel responses. The simulated voxel responses where then combined with the estimated fMRI noise using the data from the study's pilot scans. This was carried out using the *CNR/Noise_SD* approach outlined in [95]. The simulated fMRI data was then transformed into voxel RDMs and correlated with stimulus RDMs. Briefly, the steps were as follows:

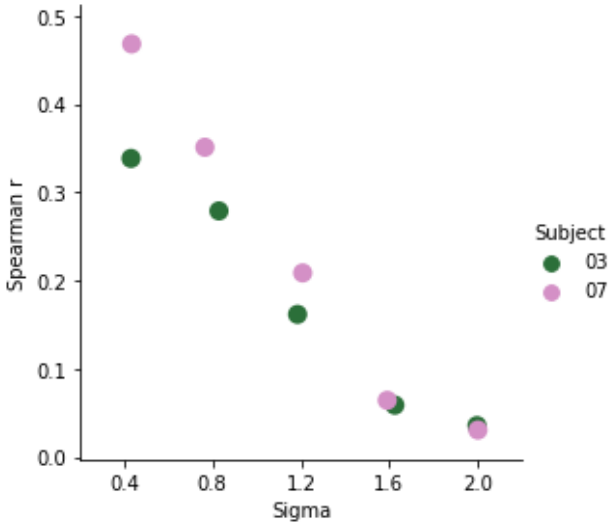
1. Estimation of population responses according to the sequence representation model given some estimate of the representational noise.
2. Estimation of fMRI noise from unprocessed EPI scans per subject.
3. Combination of simulated population responses with the estimated fMRI noise, resulting in simulated fMRI responses to individual sequences. This step was carried out using the approach and scripts developed by [96], building on previous work by [95].
4. RSA simulation: RSA carried out as described in *Methods* over simulated fRMI data to estimate a relationship between representational noise and the noise in RDMs.

The full technical details and results of the simulation are presented in the Jupyter Notebook (*sim_fmri*) at our code repository.

The simulation results are displayed in the plot below outlining the change in the RSA correlation values as a function of representational and measurement (fMRI) noise.

Our simulation shows that we can indeed expect to see a correspondence between representational noise and RSA correlation values as assumed by the SNR hypothesis: RSA correlation values decrease as noise in the sequence representations increases. The individual points om the figure above represent two different fMRI noise estimates corresponding to two subjects we scanned in the piloting phase. The difference between the first two noise parameter values ($\sigma = 0.4$ and $\sigma = 0.8$) corresponds to the estimated noise difference in the novel and learned

Fig 11: Y-axis shows the correlation between stimulus RDM and voxel RDM (Spearman's ρ) and X-axis shows the amount of noise in the representational model (as represented by the σ noise parameter).



sequence representations.

Associative learning of overlapping sequences

Worked example

To continue with the example provided in the manuscript: "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."

When two sequences ABCD and BADC are learned by strengthening item-position associations then (all other variables remaining the same) we end up with equal strengths for the following item-position associations:

$$A - 1,$$

$$A - 2,$$

$$B - 1,$$

$$B - 2,$$

$$C - 3,$$

$$C - 4,$$

$$D - 3,$$

$$D - 4.$$

The resulting weights would also be the result of learning two different sequences ABDC and BACD. In other words, learning the two original sequences would result in eight association weights of equal strength to represent four sequences (ABCD, BADC and ABDC, BACD). Such a learning mechanism would suffer from catastrophic interference with multiple short sequences of overlapping items (like most real-word sequential actions tend to be).

References

[1] Stanislas Dehaene, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, 88(1):2–19, 2015. doi: 10.1016/j.neuron.2015.09.019.

[2] James H. Howard, Darlene V. Howard, Karin C. Japikse, and Guinevere F. Eden. Dyslexics are impaired on implicit higher-order sequence learning, but not on implicit spatial context learning. *Neuropsychologia*, 44(7):1131–1144, 2006. doi: 10.1016/j.neuropsychologia.2005.10.015.

[3] Delphine Lassus-Sangosse, Marie-Ange N'guyen-Morel, and Sylviane Valdois. Sequential or simultaneous visual processing deficit in developmental dyslexia? *Vision Research*, 48 (8):979–988, 2008. doi: 10.1016/j.visres.2008.01.025.

[4] J Fiser, P Berkes, G Orbán, and M Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–30, 2010. doi: 10.1016/j.tics.2010.01.003.

[5] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–8, 2013. doi: 10.1038/nn.3495.

[6] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009. doi: 10.1016/j.cognition.2009.03.008.

[7] M Frank, S Goldwater, T Griffiths, and J Tenenbaum. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125, 2010. doi: 10.1016/j.cognition.2010.07.005.

[8] J Fiser and R N Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, 12(6):499–504, 2001.

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

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

[11] V Bejjanki, J Beck, Z Lu, and A Pouget. Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, 14(5):642–648, 2011. doi: 10.1038/nn.2796.

[12] B Lake, R Salakhutdinov, and J Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050.

[13] Konrad P. Körding and Daniel M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004. doi: 10.1038/nature02169.

[14] Emanuel Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7 (9):907–915, 2004. doi: 10.1038/nn1309.

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

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

[17] Basile Pinsard, Arnaud Boutin, Ella Gabitov, Ovidiu Lungu, Habib Benali, and Julien Doyon. Consolidation alters motor sequence-specific distributed representations. *eLife*, 8: 1–29, 2019. doi: 10.7554/elife.39324.

[18] Tobias Wiestler and Jörn Diedrichsen. Skill learning strengthens cortical representations of motor sequences. *eLife*, 2:1–20, 2013. doi: 10.7554/elife.00801.

[19] Atsushi Yokoi and Jörn Diedrichsen. Neural Organization of Hierarchical Motor Sequence Representations in the Human Neocortex. *Neuron*, pages 1–13, 2019. doi: 10.1016/j.neuron.2019.06.017.

[20] Patrick Beukema, Jörn Diedrichsen, and Timothy D. Verstynen. Binding During Sequence Learning Does Not Alter Cortical Representations of Individual Actions. *The Journal of Neuroscience*, 39(35):6968–6977, 2019. doi: 10.1523/jneurosci.2669-18.2019.

[21] Atsushi Yokoi, Spencer A. Arbuckle, and Jörn Diedrichsen. The Role of Human Primary Motor Cortex in the Production of Skilled Finger Sequences. *The Journal of Neuroscience*, 38(6):1430–1442, 2018. doi: 10.1523/jneurosci.2798-17.2017.

[22] Maxime Maheu, Stanislas Dehaene, and Florent Meyniel. Brain signatures of a multiscale process of sequence learning in humans. *eLife*, 8:1–24, 2019. doi: 10.7554/elife.41541.

[23] J. Doyon, E. Gabitov, S. Vahdat, O. Lungu, and A. Boutin. Current issues related to motor sequence learning in humans. *Current Opinion in Behavioral Sciences*, 20:89–97, 2018. doi: 10.1016/j.cobeha.2017.11.012.

[24] Nick Cumming, Mike Page, and Dennis Norris. Testing a positional model of the hebb effect. *Memory*, 11(1):43–63, 2003.

[25] A Perlman, E Pothos, D Edwards, and J Tzelgov. Task-relevant chunking in sequence learning. *Journal of experimental psychology. Human perception and performance*, 36(3): 649–61, 2010. doi: 10.1037/a0017178.

[26] Lauren K. Slone and Scott P. Johnson. When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, 178(June 2015):92–102, 2018. doi: 10.1016/j.cognition.2018.05.016.

[27] G Bower and D Winzenz. Group structure, coding, and memory for digit series. *Journal of Experimental Psychology*, 80(2, Pt.2):1–17, 1969. doi: 10.1037/h0027249.

[28] D Winzenz. Group structure and coding in serial learning. *Journal of Experimental Psychology*, 92(1):8–19, 1972. doi: 10.1037/h0032161.

[29] T Berdyyeva and C Olson. 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, 2010. doi: 10.1152/jn.00639.2009.

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

[31] T Berdyyeva and C Olson. 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, 2011. doi: 10.1152/jn.00903.2010.

[32] A Nieder, I Diester, and O Tudusciuc. Temporal and Spatial Enumeration Processes in the Primate Parietal Cortex. *Science*, 313(5792):1431–1435, 2006. doi: 10.1126/science.1130308.

[33] Ann M Graybiel, Toshihiko Aosaki, Alice W Flaherty, and Minoru Kimura. The basal ganglia and adaptive motor control. *Science*, 265(5180):1826–1831, 1994.

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

[35] Edward Ester, Thomas Sprague, and John Serences. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, 87 (4):893–905, 2015. doi: 10.1016/j.neuron.2015.07.013.

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

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

[38] Andrew C Heusser, David Poeppel, Youssef Ezzyat, and Lila Davachi. Episodic sequence memory is supported by a theta–gamma phase code. *Nature Neuroscience*, 19(10):1374–1380, 2016. doi: 10.1038/nn.4374.

[39] Liang-Tien Hsieh, Matthias J. Gruber, Lucas J. Jenkins, and Charan Ranganath. Hip-

pocampal Activity Patterns Carry Information about Objects in Temporal Context. *Neuron*, 81(5):1165–1178, 2014. doi: 10.1016/j.neuron.2014.01.015.

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

[41] D MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, 2003.

[42] Peiran Gao, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, page 214262, 2017. doi: 10.1101/214262.

[43] Stefano Fusi, Earl K. Miller, and Mattia Rigotti. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, 2016. doi: 10.1016/j.conb.2016.01.010.

[44] Lila Davachi and Sarah DuBrow. How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, 19(2):92–99, 2015. doi: 10.1016/j.tics.2014.12.004.

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

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

[47] Michael J Kahana, Matthew V Mollison, and Kelly M Addis. Positional cues in serial learning: The spin-list technique. *Memory & Cognition*, 38(1):92–101, 2010.

[48] Andrew E. Papale and Bryan M. Hooks. Circuit Changes in Motor Cortex During Motor Skill Learning. *Neuroscience*, 368:283–297, 2018. doi: 10.1016/j.neuroscience.2017.09.010.

[49] John W. Krakauer and Reza Shadmehr. Consolidation of motor memory. *Trends in Neurosciences*, 29(1):58–64, 2006. doi: 10.1016/j.tins.2005.10.003.

[50] M Page, N Cumming, D Norris, A McNeil, and G Hitch. Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, 69(4): 506–526, 2013. doi: 10.1016/j.jml.2013.07.001.

[51] D Fendrich, A Healy, and L Bourne. Long-term repetition effects for motoric and perceptual procedures. *Journal of experimental psychology. Learning, memory, and cognition*, 17 (1):137–51, 1991.

[52] K Oberauer and N Meyer. The contributions of encoding, retention, and recall to the Hebb effect. *Memory*, 17(7):774–81, 2009. doi: 10.1080/09658210903107861.

[53] Kristjan Kalm and Dennis Norris. Recall is not necessary for verbal sequence learning. *Memory & cognition*, 44(1):104–13, 2016. doi: 10.3758/s13421-015-0544-0.

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

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

[56] Raymond P. Kesner and Edmund T. Rolls. A computational theory of hippocampal function, and tests of the theory: New developments. *Neuroscience and Biobehavioral Reviews*, 48:92–147, 2015. doi: 10.1016/j.neubiorev.2014.11.009.

[57] N White and R McDonald. Multiple parallel memory systems in the brain of the rat. *Neurobiology of learning and memory*, 77(2):125–84, 2002. doi: 10.1006/nlme.2001.4008.

[58] N Voermans, K Petersson, L Daudey, B Weber, K Van Spaendonck, H Kremer, and Guillén Fernández. Interaction between the human hippocampus and the caudate nucleus during route recognition. *Neuron*, 43(3):427–35, 2004. doi: 10.1016/j.neuron.2004.07.009.

[59] J L McClelland and N H Goddard. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6(6):654–65, 1996. doi: 10.1002/(sici)1098-1063(1996)6:6¡654::aid-hipo8¿3.0.co;2-g.

[60] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 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, 1995. doi: 10.1037/0033-295x.102.3.419.

[61] M. Bunsey and H. Eichenbaum. Conservation of hippocampal memory function in rats and humans. *Nature*, 379(6562):255–257, 1996. doi: 10.1038/379255a0.

[62] B Murdock. Context and mediators in a theory of distributed associative memory (TODAM2)., 1997.

[63] Gordon DA Brown, Tim Preece, and Charles Hulme. Oscillator-based memory for serial order. *Psychological Review*, 107(1):127, 2000.

[64] Neil Burgess and Graham J Hitch. Toward a network model of the articulatory loop. *Journal of Memory and Language*, 31(4):429–460, 1992.

[65] Michael E Hasselmo, Marc W Howard, Mrigankka Fotedar, and Aditya V Datey. 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, 2006.

[66] C Lee and W Estes. 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, 1981.

[67] R Henson. Positional information in short-term memory: relative or absolute? *Memory & Cognition*, pages 915–927, 1999.

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

[69] R Henson and N Burgess. Representations of Serial Order. *The Quarterly Journal of Experimental Psychology*, 1997.

[70] G Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63:81–97, 1956.

[71] Steven J. Luck and Edward K. Vogel. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, 17(8):391–400, 2013. doi: 10.1016/j.tics.2013.06.006.

[72] R van den Berg, E Awh, and W Ma. Factorial comparison of working memory models. *Psychological review*, 121(1):124–49, 2014. doi: 10.1037/a0035234.

[73] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4), 2014. doi: 10.1371/journal.pcbi.1003553.

[74] P Schramm and J Rouder. Are Reaction Time Transformations Really Beneficial? *PsyArXiv*, 2019. doi: 10.31234/osf.io/9ksa6.

[75] Harald Baayen and Petar Milin. Analyzing reaction times. *International Journal of Psychological Research*, 3(2):12, 2010. doi: 10.21500/20112084.807.

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

[77] Oscar Esteban, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit Ghosh, Jessey Wright, Joke Durnez, Russell Poldrack, and Krzysztof Jacek Gorgolewski. Fmriprep: a robust preprocessing pipeline for functional MRI. *bioRxiv*, 2018. doi: 10.1101/306951.

[78] Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, Mathias Kent, James D. and-Goncalves, Elizabeth DuPre, Kevin R. Sitek, Daniel E. P. Gomez, Daniel J. Lurie, Zhifang Ye, Russell Poldrack, and Krzysztof Jacek Gorgolewski. Fmriprep 1.1.7. *Software*, 2018. doi: 10.5281/zenodo.852659.

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

[80] Krzysztof Jacek Gorgolewski, Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, Hans Johnson, Christopher Burns, Alexandre Manhães-Savio, Carlo Hamalainen, Benjamin Yvernault, Taylor Salo, Kesshi Jordan, Mathias Goncalves, Michael Waskom, Daniel Clark, Jason Wong, Fred Loney, Marc Modat, Blake E Dewey, Cindee Madison, Matteo Visconti di Oleggio Castello, Michael G. Clark, Michael Dayan, Dav Clark, Anisha Keshavan, Basile Pinsard, Alexandre Gramfort, Shoshana Berleant, Dylan M. Nielson, Salma Bougacha, Gael Varoquaux, Ben Cipollini, Ross Markello, Ariel Rokem, Brendan Moloney, Yaroslav O. Halchenko, Demian Wassermann, Michael Hanke, Christian Horea, Jakub Kaczmarzyk, Gilles de Hollander, Elizabeth DuPre, Ashley Gillman, David Mordom, Colin Buchanan, Rosalia Tungaraza, Wolfgang M. Pauli, Shariq Iqbal, Sharad Sikka, Matteo Mancini, Yannick Schwartz, Ian B. Malone, Mathieu Dubois, Caroline Frohlich, David Welch, Jessica Forbes, James Kent, Aimi Watanabe, Chad Cumba, Julia M. Huntenburg, Erik Kastman, B. Nolan Nichols, Arman Eshaghi, Daniel Ginsburg, Alexander Schaefer, Benjamin Acland, Steven Giavasis, Jens Kleesiek, Drew Erickson, René Küttner, Christian Haselgrove, Carlos Correa, Ali Ghayoor, Franz Liem, Jarrod Millman, Daniel Haehn, Jeff Lai, Dale Zhou, Ross Blair, Tristan Glatard, Mandy Renfro, Siqi Liu, Ari E. Kahn, Fernando Pérez-García, William Triplett, Leonie Lampe, Jörg Stadler, Xiang-Zhen Kong, Michael Hallquist, An-

drey Chetverikov, John Salvatore, Anne Park, Russell Poldrack, R. Cameron Craddock, Souheil Inati, Oliver Hinds, Gavin Cooper, L. Nathan Perkins, Ana Marina, Aaron Mattfeld, Maxime Noel, Lukas Snoek, K Matsubara, Brian Cheung, Simon Rothmei, Sebastian Urchs, Joke Durnez, Fred Mertz, Daniel Geisler, Andrew Floren, Stephan Gerhard, Paul Sharp, Miguel Molina-Romero, Alejandro Weinstein, William Broderick, Victor Saase, Sami Kristian Andberg, Robbert Harms, Kai Schlamp, Jaime Arias, Dimitri Papadopoulos Orfanos, Claire Tarbert, Arielle Tambini, Alejandro De La Vega, Thomas Nickson, Matthew Brett, Marcel Falkiewicz, Kornelius Podranski, Janosch Linkersdörfer, Guillaume Flandin, Eduard Ort, Dmitry Shachnev, Daniel McNamee, Andrew Davison, Jan Varada, Isaac Schwabacher, John Pellman, Martin Perez-Guevara, Ranjeet Khanuja, Nicolas Pannetier, Conor McDermottroe, and Satrajit Ghosh. Nipype. *Software*, 2018. doi: 10.5281/zenodo.596855.

[81] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6): 1310–1320, 2010. doi: 10.1109/tmi.2010.2046908.

[82] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. doi: 10.1006/nimg.1998.0395.

[83] Arno Klein, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, Brian Rossa, Martin Reuter, Elias Chaibub Neto, and Anisha Keshavan. Mind-boggling morphometry of human brains. *PLOS Computational Biology*, 13(2):e1005350, 2017. doi: 10.1371/journal.pcbi.1005350.

[84] VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1:S102, 2009. doi: 10.1016/s1053-8119(09)70884-5.

[85] B.B. Avants, C.L. Epstein, M. Grossman, and J.C. Gee. Symmetric diffeomor-

phic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008. doi: 10.1016/j.media.2007.06.004.

[86] Y. Zhang, M. Brady, and S. Smith. 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, 2001. doi: 10.1109/42.906424.

[87] Douglas N Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009. doi: 10.1016/j.neuroimage.2009.06.060.

[88] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002. doi: 10.1006/nimg.2002.1132.

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

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

[91] Jeanette Mumford, Benjamin Turner, Gregory Ashby, and Russell Poldrack. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–43, 2012. doi: 10.1016/j.neuroimage.2011.08.076.

[92] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006. doi: 10.1016/j.neuroimage.2006.01.021.

[93] Chih-chung Chang and Chih-jen Lin. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:1–39, 2013. doi: 10.1145/1961189.1961199.

[94] Jörn Diedrichsen, Serge Provost, and Hossein Zareamoghaddam. On the distribution of cross-validated Mahalanobis distances. *arXiv*, pages 1–24, 2016.

[95] Marijke Welvaert and Yves Rosseel. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. *PLoS ONE*, 8(11), 2013. doi: 10.1371/journal.pone.0077089.

[96] Cameron T Ellis, Christopher Baldassano, Anna C Schapiro, Ming Bo Cai, and Jonathan D Cohen. Facilitating open-science with realistic fMRI simulation: validation and application. *bioRxiv*, 2019. doi: 10.1101/532424.