J Neurophysiol 112: 1849–1856, 2014.
First published July 30, 2014; doi:10.1152/jn.00028.2014.

Multifaceted aspects of chunking enable robust algorithms

Daniel E. Acuna,[1] Nicholas F. Wymbs,[2] Chelsea A. Reynolds,[3] Nathalie Picard,[4] Robert S. Turner,[4] Peter L. Strick,[4] Scott T. Grafton,[2] and Konrad P. Kording[1]

[1]*Rehabilitation Institute of Chicago and Northwestern University, Chicago, Illinois;* [2]*University of California, Santa Barbara, California;* [3]*Center for Neuroscience and the Center for the Neural Basis of Cognition, University of Pittsburg, Pittsburgh, Pennsylvania; and* [4]*University of Pittsburgh, Pittsburgh, Pennsylvania*

Submitted 9 January 2014; accepted in final form 21 July 2014

Acuna DE, Wymbs NF, Reynolds CA, Picard N, Turner RS, Strick PL, Grafton ST, Kording KP. Multifaceted aspects of chunking enable robust algorithms. *J Neurophysiol* 112: 1849–1856, 2014. First published July 30, 2014; doi:10.1152/jn.00028.2014.—Sequence production tasks are a standard tool to analyze motor learning, consolidation, and habituation. As sequences are learned, movements are typically grouped into subsets or chunks. For example, most Americans memorize telephone numbers in two chunks of three digits, and one chunk of four. Studies generally use response times or error rates to estimate how subjects chunk, and these estimates are often related to physiological data. Here we show that chunking is simultaneously reflected in reaction times, errors, and their correlations. This multimodal structure enables us to propose a Bayesian algorithm that better estimates chunks while avoiding overfitting. Our algorithm reveals previously unknown behavioral structure, such as an increased error correlations with training, and promises a useful tool for the characterization of many forms of sequential motor behavior.

discrete sequence production; learning; memory

ONE OF THE CENTRAL QUESTIONS in neuroscience and cognitive and behavioral psychology is how the brain allows movement performance to improve with practice (Thorndike 1898; Pavlov 1927; Crossman 1959; Newell and Rosenbloom 1981; Skinner 1938). When learning to produce sequences of actions, animals organize information into groups or "chunks" (Miller and George 1956; Newell 1991), and studying such chunks is a popular way of studying learning and memory. One common experimental paradigm to study movement chunking is the discrete sequence production (DSP) task, where subjects learn, through many repetitions, to rapidly generate a fixed sequence of finger movements (Adams 1984; Gentner 1987; Verwey 1996; Verwey and Dronkert 1996; Logan and Bundesen 2003). Understanding how practice leads to improved performance in this task promises to unveil how the central nervous system organizes temporal information in a way that enables fast, habitual processing (e.g., see Clerget et al. 2012).

Indeed, evidence suggests that chunking is what allows efficient behavior. Performance gains in the DSP task correlate with motor chunking (Verwey 1994, 1996; Verwey and Dronkert 1996). This may be because long sequences cannot be stored in short-term motor memory or the optimal control problem for long sequences is generally infeasible (Todorov et al. 2005; Parr 1998). Understanding motor chunking is crucial to understanding the organization of motor memory and movement efficiency.

Motor chunks are known to leave characteristic traces on observed response times. One of the prominent features is a change in response times and errors at the beginning of chunks. There is a longer-than-usual pause during chunk concatenation (Verwey and Dronkert 1996; Verwey 1996). These differences in response times are frequently described and used in algorithms that reveal chunking behavior.

It seems natural to think of each chunk as controlled by a single neural representation, and each representation should be able to produce these movements at the right speed, leading to additional predictions of features of chunking. When a new chunk is started, the transition should slow response times, and errors in the transition should lead to errors in the execution. However, we should also expect that each chunk is produced at a relatively fixed speed (Abrahamse et al. 2013). Reaction times and errors within a chunk should therefore be correlated, forming a novel hypothesis to test.

There exists a range of methods for inferring chunking structure. The most common is to look at the mean response times during a set of trials and detect significant increases in certain elements of the sequence (Abrahamse et al. 2013; Verwey 1994, 1996; Verwey and Dronkert 1996). These points of significant increase are marked as the start of a new chunk. This simple method limits the analysis by requiring the experimenter to choose which ranges of trials to analyze. A recent approach based on community detection in networks (Wymbs et al. 2012) removes this limitation by modeling time explicitly and providing time-varying evolution of chunk structures. These methods, however, cannot readily incoporate multiple signals (response times and errors) or deal with correlations of these signals within trials.

In this article, we analyze data from 17 subjects who participated in a DSP task over an average of 30 days of practice. We do find that the features reaction time, error, and their respective correlations are associated with chunking structure. Combining information across features and time allows us to construct a Bayesian algorithm that is consistently more precise than algorithms using only one feature. To verify the general applicability of our algorithm, we also analyze two datasets from nonhuman primates producing and learning sequences over months or even years. The resulting algorithm that combines the multimodal features of behavior and across time allows precise estimates of underlying chunking structure.

Address for reprint requests and other correspondence: D. E. Acuna, Rehabilitation Institute of Chicago and Northwestern Univ., Chicago, IL (e-mail: daniel.acuna@northwestern.edu).

METHODS

Experimental Data

We analyze data where humans or monkeys participate in a DSP task. Subjects observe a cue that signals the next element of a sequence of elements to be executed. Each element of the sequence is signaled to the subject one at a time and after inputting the right element, the system advances to the next element, and so on until the end of the sequence (see Fig. 2*A*), which allows subjects to speed up the task by predicting the next element.

Human data. We reanalyze data from a published study (Bassett et al. 2013). In short, 25 subjects completed a training regimen involving the simultaneous acquisition of 6 different 10-element motor sequences using a DSP task. Subjects were asked to input the elements as fast and accurate as possible. Over the course of a 6-wk training regimen, subjects trained on the sequences both at home using their laptop computer as well as inside an MRI scanner. Training alternated between scanner and home locations, such that following each training session in the scanner, subjects performed a minimum of 10 sessions (1 session/day) at home over the next 2 wk. This pattern of training repeated 3 times, so that subjects completed on average 30 home training sessions and 4 scan sessions. We only analyze the home training sessions. Sequence familiarity was manipulated during home training at three exposure levels. Each home training session consisted of 150 trials presented using a random schedule, so that two sequences trained extensively (~2,000 trials), two sequences trained moderately, and two sequences trained minimally. All subjects trained on the same sequence set and each at the same exposure level, which were maintained over the course of training. Only the highly trained sequences were analyzed in the present study. Sequence one is 2, 1, 3, 5, 4, 1, 3, 4, 5, 2, and 5, and sequence two is 4, 2, 1, 3, 5, 2, 3, 1, 4, and 5. Fingers are sequentially numbered from (1) the thumb to (5) the pinky. We excluded eight subjects because of either lower participation rate, increase response times, or error rates with practice.

Monkey data. We reanalyze data from a published study (Matsuzaka et al. 2007, dataset D1): in short, two rhesus macaques participated in a 12-element 5-target DSP task. We analyze data from one monkey. Targets for the reaching movements were on a touch screen, chosen by touching the screen with the hand, and visually cued 300 ms after the previous target contact. However, the task allowed the monkeys to contact the next target in the sequence during the 300-ms delay before it was shown. When the monkeys made correct anticipatory responses, the task was incremented to the next element of the sequence without display of the touched target. As a result, the monkeys learned to perform the sequence in a predictive fashion without requiring visual cues.

We reanalyze a second dataset from a previous study (Desmurget and Turner 2010, dataset D2): in this task, two rhesus macaques participated in a five-element eight-target DSP task (see also Desmurget and Turner 2008, 2010). Targets for the movements were displayed on a monitor, chosen by steering a cursor through a target area with a joystick, and visually cued 230 ms after the previous target acquisition. The task required the animal to move the cursor in five consecutive out and back movements between the central start position and a series of five peripheral targets. As in dataset D1, with training, the monkey performed the task in a predictive fashion without relying on visual cues.

Preprocessing

The aim of this study is to understand the evolution of chunking structure over time. For this purpose, it is necessary to control for the effect of practice on overall movement speed that is unrelated to the within-sequence variability introduced by chunking. For example, by computing the mean response time as a function of trials, we can clearly see a trend that comes from practicing (see Fig. 2*B*). In our analysis, we remove this unrelated effect by performing a detrending of the response time data using an exponential model (Logan and Bundesen 2003). There are alternative models for detrending practice curves, such as the power law, and depending on the circumstance different curves might be used (Heathcote et al. 2000).

We assume that the practice component of response time depends on trials across days and the interaction between elements and trial

$$\mathrm{RT}_{t,i} = a_0 + \underbrace{a_1 \cdot \exp\left(-b_1 \cdot t - b_2 \cdot t \cdot i\right)}_{\text{Practice effect}} + \underbrace{\mathrm{RRT}_{t,i}}_{\text{Chunking+noise}}, \tag{1}$$

where t is the trial and i is the element. We control for the interaction of trial and element because response time tends to be faster in later elements of the sequence, independent of the chunking structure (Abrahamse et al. 2013). Other covariates assumed to be independent of the chunking structure could be included in the detrending preprocesing, such as finger used

$$\mathrm{RT}_{t,i} = a_0 + \underbrace{a_1 \cdot \exp\left(-b_1 \cdot t - b_2 \cdot t \cdot i\right)}_{\text{Practice effect}} + \underbrace{\sum_{k=2}^{5} \mathrm{finger}_k}_{\text{Biomechanical effect}} + \underbrace{\mathrm{RRT}_{t,i}}_{\text{Chunking+noise}} \tag{2}$$

The residual response time (RRT), then, is assumed to come from the joint contribution of chunking structure and the motor variability (noise). This allows us to use the reaction times as a meaningful way of estimating chunking structure.

Generative Model for RRT and Errors

Our Bayesian algorithm is based on a so-called generative model, a probabilistic description of how statistical properties of the task, the chunking structure, are reflected in the statistical properties of measurable data-response time and errors. In our modeling, "error" refers to the failure of a participant to correctly generate an element and it is therefore an observation of the behavior rather than a *statistical* estimation error. Here we assume that the chunking structure tends to stay similar from trial to trial but may change a great deal over many trials. Based on preliminary analysis described below, we build into the model the assumption that the first element of each chunk may have prolonged RRT and raised error probabilities and that, within each chunk, response times and error locations are correlated.

We thus can define a probabilistic graphical model (Jordan 2004) of how the data are stochastically related within a trial and across trials. The RRTs depend on the parameters of the features and the chunking structure at trial *t*. The goal of our algorithm is to estimate these parameters.

We assume that there is a chunking structure, c_t, which slowly evolves across trials. Statistically, this is captured by the following generative model:

$$c_0 \sim \pi, \tag{3}$$

$$c_t \big| c_{t-1} \sim \Pi\left(c_{t-1} \to \cdot\right), \tag{4}$$

$$\mathrm{RRT}_t \big| c_t \sim \mathrm{Normal}\left(\mu_{c_t}^{\mathrm{RRT}}, \sum\nolimits_{c_t}^{\mathrm{RRT}}\right), \tag{5}$$

$$\mathrm{ER}_t \big| c_t \sim \mathrm{Normal}\left(\mu_{c_t}^{\mathrm{ER}}, \sum\nolimits_{c_t}^{\mathrm{ER}}\right). \tag{6}$$

Our model is a hidden Markov model with Gaussian emissions (Murphy 2012). There is a finite number of chunking structures ($c \in \{1, ..., K\}$), and the matrix $\Pi(c \to c')$ governs the probability of going from chunking structure c to c'. The distribution π is the probability of starting at any of the K chunking structures. Conditioned on the chunking structure, multivariate normal distributions are assumed for RRT and errors (*Eqs. 5* and *6*). Modeling errors as normally distributed can only capture their means and covariance but not their skewness, but we made this modeling choice for statistical simplicity and computational efficiency. More specifically, the mean RRT at trial t for element i is

Table 1. *Parameters needed by the model to estimate chunks*

Parameter	Description
π	Initial distribution of chunking structures
Π	Transition probability matrix between chunking structures
$\mu^{\mathrm{RRT}}_{\mathrm{start}}$	Mean RRT at beginning of chunk
$\mu^{\mathrm{RRT}}_{\mathrm{nonstart}}$	Mean RRT within a chunk
$\mu^{\mathrm{ER}}_{\mathrm{start}}$	Error rate at beginning of chunk
$\mu^{\mathrm{ER}}_{\mathrm{nonstart}}$	Error rate within a chunk
σ^2_{RRT}	RRT variance
σ^2_{ER}	Error rate variance
ρ_{RRT}	Correlation between RRT of elements within same chunk
ρ_{ER}	Correlation between error of elements within same chunk

These parameters are fitted by expectation maximization. RRT, residual response time.

$$E[\mathrm{RRT}_{t,i}|c_t] = \begin{cases} \mu^{\mathrm{RRT}}_{\mathrm{start}} & i \text{ is start of chunk} \\ \mu^{\mathrm{RRT}}_{\mathrm{nonstart}} & \text{otherwise} \end{cases} \tag{7}$$

and, similarly, the mean error at trial t for element i is

$$E[\mathrm{ER}_{t,i}|c_t] = \begin{cases} \mu^{\mathrm{ER}}_{\mathrm{start}} & i \text{ is start of chunk} \\ \mu^{\mathrm{ER}}_{\mathrm{nonstart}} & \text{otherwise} \end{cases} \tag{8}$$

The variance across elements is assumed to be homogenous over chunks, i.e., $\mathrm{Var}[\mathrm{RRT}_{t,i}|c_t] = \sigma^2_{\mathrm{RRT}}$ and $\mathrm{Var}[\mathrm{ER}_{t,i}|c_t] = \sigma^2_{\mathrm{ER}}$.

The covariance takes on a simple structure, assumed constant within a chunk and zero across chunks. Because a constant variance is assumed as well, the correlation within a chunk is constant, and the correlation across chunks is null. Mathematically, the correlations in RRTs and errors at trial t between elements i and j are

$$\mathrm{Corr}[\mathrm{RRT}_{t,i}, \mathrm{RRT}_{t,j}|c_i] = \begin{cases} \rho_{\mathrm{RRT}} & i \text{ and } j \text{ are in same chunk} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and

$$\mathrm{Corr}[\mathrm{ER}_{t,i}, \mathrm{ER}_{t,j}|c_i] = \begin{cases} \rho_{\mathrm{ER}} & i \text{ and } j \text{ are in same chunk} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

To infer the chunking structure, we must simultaneously learn the parameters π, Π, $\mu^{\mathrm{RRT}}_{\mathrm{start}}$, $\mu^{\mathrm{RRT}}_{\mathrm{nonstart}}$, $\mu^{\mathrm{ER}}_{\mathrm{start}}$, $\mu^{\mathrm{ER}}_{\mathrm{nonstart}}$, σ^2_{RRT}, σ^2_{ER}, ρ_{RRT}, and ρ_{ER} and compute the distribution over chunking structures (see Table 1). We use a standard expectation maximization scheme in which we alternate between finding the expected chunking structure of the model through a forward-backward algorithm (Welch 2003) and then finding the parameters that most likely would have generated those chunking structures. This scheme is proven to find a local maximum on log-likelihood (Murphy 2012). The algorithm optimally estimates an interpretation of the data (response times, errors, and correlations) in terms of a slowly varying chunking structure. The code of the algorithm is available at https://github.com/daniel-acuna/chunk_inference).

Generation of Candidate Chunking Structures

A candidate set of chunking structures must be provided to our method. In this study, we tested all combinations of chunking structures with one or more elements. For example, for a 10-element sequence, there are 511 possible chunking structures, from the hypothetical that each element is in its own chunk to all 10 elements being in one chunk (Fig. 1*A*). Each chunking structure defines mean and

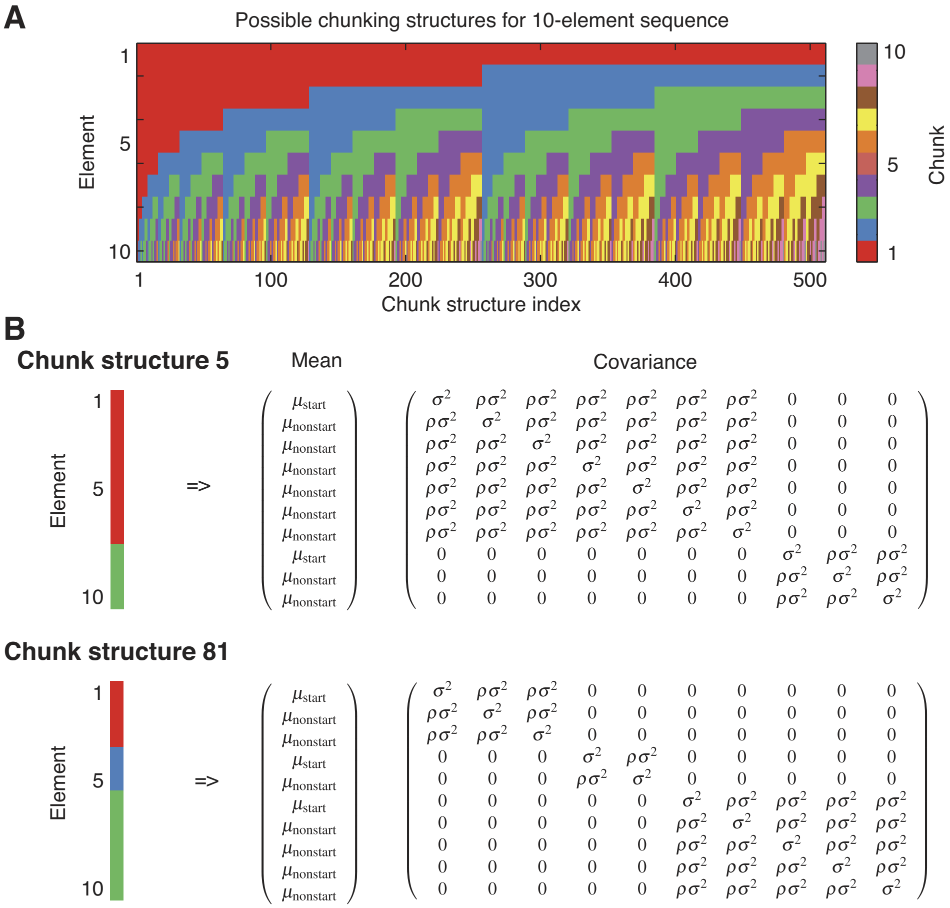

Fig. 1. *A*: exhaustive list of chunking structures that might be present during a trial of a 10-element sequence. *B*: example of mean and covariance of residual response time (RRT) or error of the elements given 2 example chunking structures.

covariance RRTs and errors, as defined by *Eqs. 7*, *8*, *9*, and *10* (for concrete examples, see Fig. 1*B*). After the list of possible chunking structures is defined, the algorithm estimates the parameters of the hidden Markov model, inferring the distribution of chunking structure present at each trial.

Alternatives to the Full Model and Model Comparison Measures

The model presented in the previous sections uses four features of the data to infer chunking structure. These features are mean RRT and errors, and RRT/RRT, and error/error correlations. By enabling or disabling each of these features, it is possible to fit simpler variations of the full model. In this study, we test three variations of the model:

1) RT + ER: full model with correlations set to zero ($\rho_{\mathrm{RRT}} = 0$, $\rho_{\mathrm{ER}} = 0$).

2) RT + ER + RRT/RRT: full model with error correlations set to zero ($\rho_{\mathrm{ER}} = 0$).

3) RT + ER + ER/ER: full model with RRT correlations set to zero ($\rho_{\mathrm{RRT}} = 0$).

Model comparison will be performed by fitting the parameters needed for each model in a *training dataset* and reporting the prediction performance of the model in a separate *testing dataset*. This procedure is a standard technique in machine learning where models of differing complexity need to be compared (Hastie et al. 2009). Specifically, if a model is too complex for a dataset, it fits the training data very well but it performs poorly during the prediction of the testing dataset due to overfitting. Similarly, if a model is excessively simple, it poorly fits the training and testing datasets. The model comparison method used here will favor models that capture real trends in the data rather than the noise.

RESULTS

We want to understand how chunking affects behavior and, based on that knowledge, build an algorithm that estimates chunks given response times, errors, and correlations. We base our analysis on results from several previously published experiments that used humans and monkeys as subjects. In the human study, subjects had to type out sequences, while given visual information about the correct key to press in sequential order (Fig. 2*A*). Similarly, in the monkey studies, subjects had to reach over a screen or use a joystick to select the correct element in the sequence. Over the course of extensive training, subjects learned to predict each element, and, by the end of the training, response times were very fast, indicating predictive behavior (~150 ms). We characterize the statistics of the multimodal structure of this data and use it to calibrate an algorithm that efficiently estimates chunking structure.

In the human dataset, subjects get better over time (Fig. 2*B*), independent of the actual chunking structure. The exponential fit (1) reveals that parameters across subjects and sequences were similar, suggesting that practice had a similar effect on response time across subjects and sequences. Subjects clearly learn to become more efficient, but this effect is orthogonal to the way they chunk the sequence.

The raw data show interesting structure that is consistent with the existence of chunks (Fig. 2*C*). We clearly see that some response times tend to be longer throughout the experiment (e.g., elements 1, 5, and 6) while others are only longer during parts of the experiment (e.g., element 7 from trial 400 to 1,400). However, chunking structure based on response time is salient early in learning and decays in importance later. Errors also follow a similar but substantially more subtle trend. Importantly, the increased error probabilities and reaction times are temporally coherent: if the time is longer on one trial, it tends to be longer on the next trial. These are signs of chunking structure and its slow evolution over time.

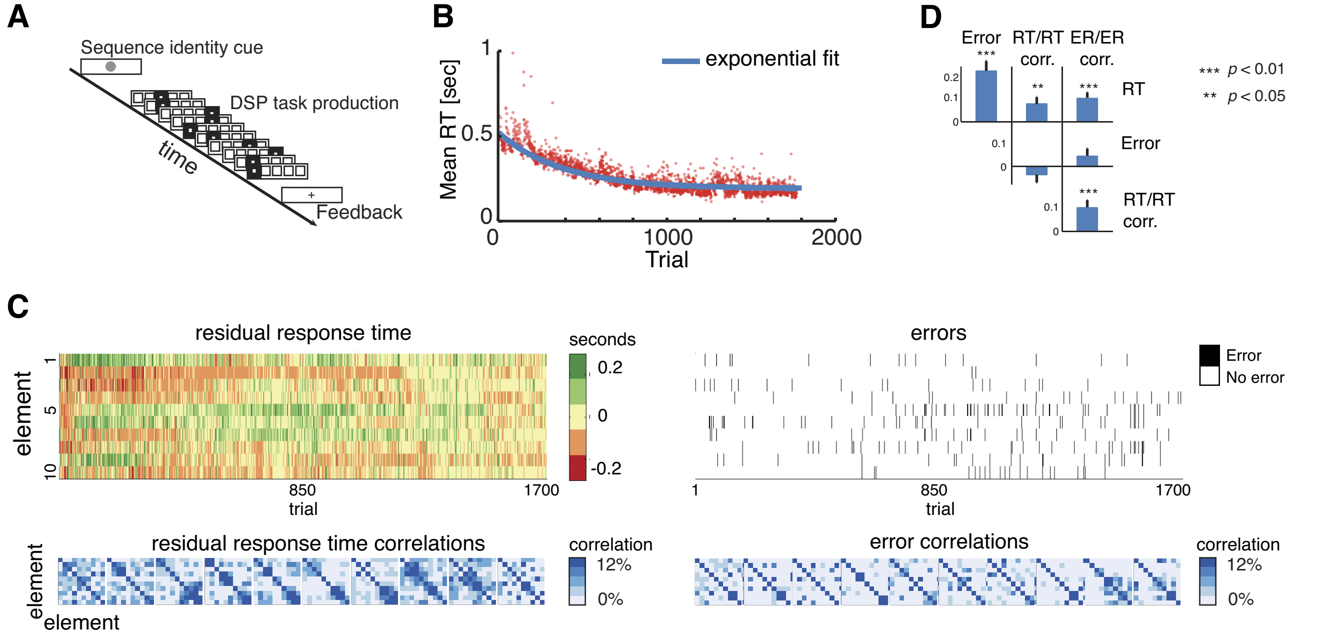

Fig. 2. *A*: experimental setup for human experiment. Subjects repeatedly and rapidly inputted a 10-element sequence of 5 targets (1 for each finger). *B*: reaction times decreased over trials and an exponential fit captures the bulk of this effect. *C*, *top left*: RRTs as a function of element (*y*-axis) and trial (*x*-axis). Green depicts a higher-than-average residual, and red depicts a lower-than-average residual. *C*, *bottom left*: element-element RRTs correlations in 200-trial blocks. Blue color represents strength of correlation. *C*, *top right*: errors by element and trials. *C*, *bottom right*: error-error correlations in 200-trial blocks. *D*: correlations between estimated features used by the algorithm. This analysis empirically shows that most features are positively associated with one another. DSP, discrete sequence production; ER, error rate.

As we hypothesized, there may be structure in the data that goes beyond mean RRTs and error probabilities. We find clear evidence of chunking structure in the RRT/RRT and error/error correlations for this example subject (Fig. 2*C*, *bottom*). Empirically, we found that these four features (mean RRT and error, and RRT/RRT, and error/error correlations) are positively associated with one another, suggesting that they represent one underlying structure. Reaction times are correlated to the probability of making an error. However, if long reaction times indicate the start of a chunk, then the next two items will usually be part of the same chunk and should have correlated reaction times. Also, as predicted, we find a correlation of reaction times with the correlation of reaction times with the next item. In addition, in the same manner we find the expected relation to the error-error correlations. Similarly, we find the correlation of errors and error-error correlations (see Fig. 2*D* for all associations across these 4 features). Chunking is not just a phenomenon of reaction times but is reflected in the full set of hypothesized features.

Our model is now driven by the insights from the data (Fig. 2*D*). *1*) Reaction times and errors, as well as their correlations across elements are indicators of chunking structure. *2*) Chunking changes slowly over time. These four features we use are complimentary: response times are powerful features early in learning but later they require significant integration over time while errors and correlational information provide a constant secondary source of information, in particular later in learning. Combining these insights allows our model to estimate the chunking structure.

Do response times and errors inform us about the same underlying chunking structure? To ask this question we ask how the models based on only response time or errors predicts the other. We find that when estimating chunking based on errors only, response times at chunk boundaries are signifi-

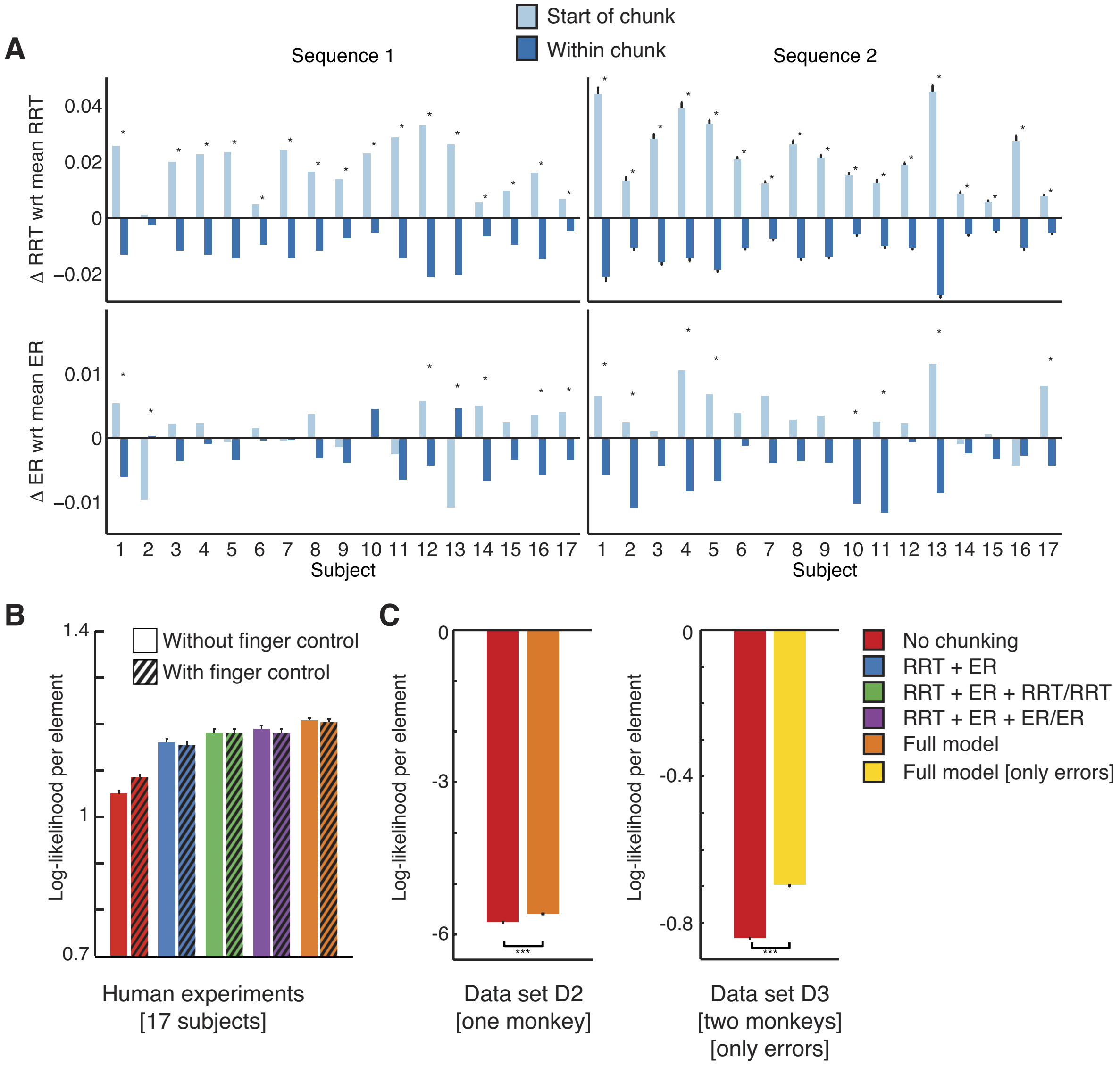

Fig. 3. Predictive quality of the model. Error bars represent 95% confidence intervals. *A*, *top*: when only errors are used to estimate chunking then chunk starts are slower. *A*, *bottom*: when only response times are used to estimate chunking then chunk starts have more errors. Error and response time data are coherent, coinciding with the algorithm's assumptions. *B* and *C*: model performance on testing data which controls for the differing complexity of the models. We compared models based on various subsets of features (see plot legends). For the human dataset, we tested the model in residual response without finger control (*Eq. 1*) and with finger control (*Eq. 2*). Across all datasets, our full model based on all features performed best without overfitting.

cantly longer than within chunks (Fig. 3*A*, *top*). We also find that when estimating chunking based on reaction times only, chunk starts have more errors (Fig. 3*A*, *bottom*). This suggests that the various features are all reflections of one underlying chunk structure.

We additionally tested the applicability of these ideas to very long learning studies. We use two datasets from two laboratories that involved macaques performing DSP tasks. The steps used to analyze the data were analogous to those used in the human experiment previously described. For one dataset (D1), from Matsuzaka et al. (2007), we analyzed response times and errors, and for the second dataset (D2), from Desmurget and Turner (2010), we only analyzed errors. Dataset D1 contains ~120,000 trials of 12-element 5-target sequences from 1 monkey, and Dataset D2 totals ~138,000 trials of 5-element 8-target sequences from 2 monkeys. The extent and controllability achieved by these experiments make them ideal for testing the introduced algorithm. Indeed, we find that for all datasets, when we train on the first 80% of the data to predict the last 20% of the data, we are better than a no-chunking algorithm (Fig. 3, *B* and *C*). These results show that our algorithm is able to describe meaningful aspects of sequence production tasks over long time scales.

A more detailed analysis of the human data with the full model is performed next. The algorithm provides us with time-varying fits to response times, errors, and correlations. The model can provide the most probable chunking structure at each trial (see example inference in Fig. 4*B*) and qualitatively fits the response times and errors, exhibiting a block-diagonal correlational structure consistent with the data (compare Fig. 2*C* with Fig. 4*C*).

Over the course of learning, the nervous system may use chunks of varying size. Indeed, we see that smaller chunks are replaced by larger ones (Fig. 5*A*), consistent with the literature (Abrahamse et al. 2013). It seems that subjects encode sequences by using many chunks at the beginning of learning and then slowly "compiling" them into longer chunks as has been suggested by the prior literature (Newell 1991; Ericsson 2006; Wymbs et al. 2012).

The algorithm uses whichever features are indicative about the chunking structure which may evolve over time. Response times and errors at chunking boundaries evolves over time (Fig. 5*B*). Response time/response time correlations remain relatively constant throughout the experiment and, interestingly, error-error correlations increase over time (Fig. 5*C*). Across learning our algorithm identifies chunks using reaction times, errors, and their correlations, and how chunks are expressed evolves over time.

DISCUSSION

In this work, we have analyzed reaction times, errors, and their correlations and found all of them to be indicative of an underlying chunking structure. Interestingly, signs of chunking structure change over time. The importance of the response time signal is larger than errors, and error correlations become increasingly important at the end of the trials. We also found chunking structure to evolve slowly over time, with more chunks at the beginning and fewer chunks at the end. Combining these properties into a joint model for chunking inference enabled an algorithm that uses the breadth of available data. Our algorithm consistently outperforms simpler algorithms across datasets from humans and monkeys. Our algorithm

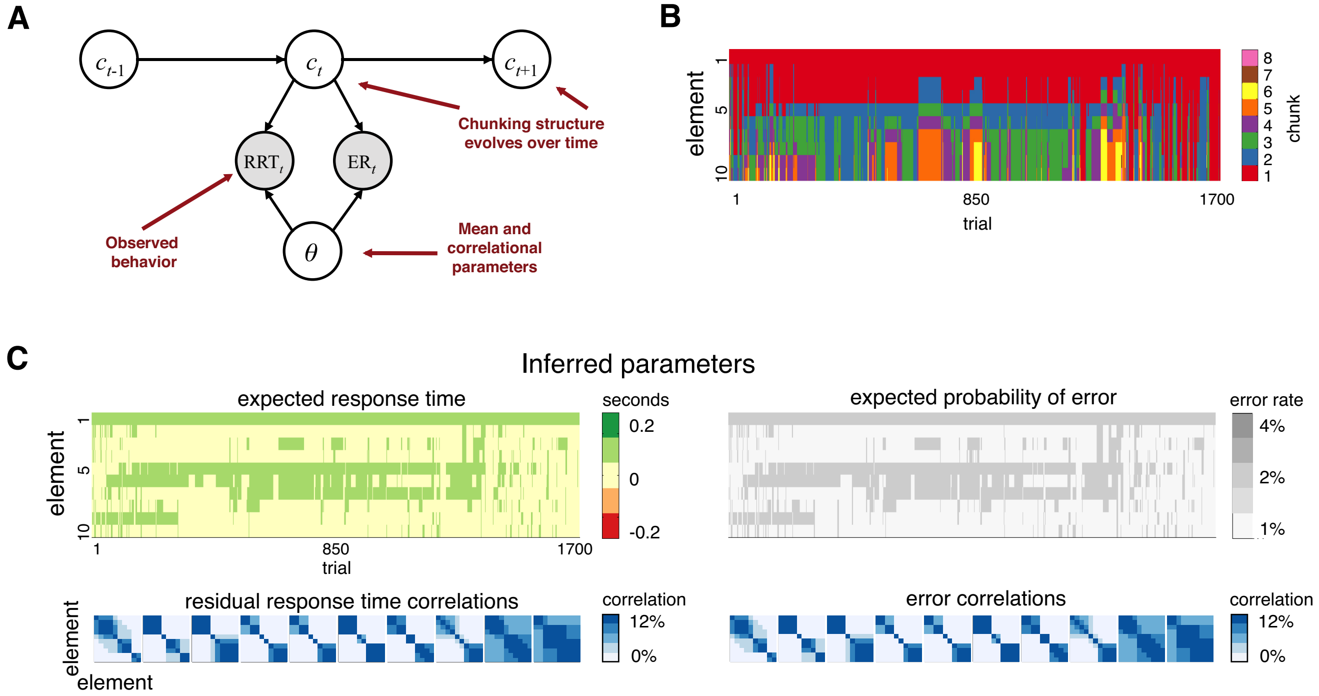

Fig. 4. *A*: chunking structure at time *t*, denoted by the node c_t, is assumed to evolve slowly over time and governs the probability of the next chunking structure c_{t+1}. Given the current chunking structure, c_t, the model assumes a likelihood on the estimated RRT. Multifaceted aspects of chunking enable robust algorithms 18 (RRT$_t$), errors (ER$_t$), and correlations (not shown). The effect of chunking structure on the data is parameterized by the vector θ. *B*: maximum a posteriori estimation of chunking structure based on the data. Each color represents a different chunk. *C*: expected mean response times and errors and their correlations inferred by the model. Compare to Fig. 2*C*.

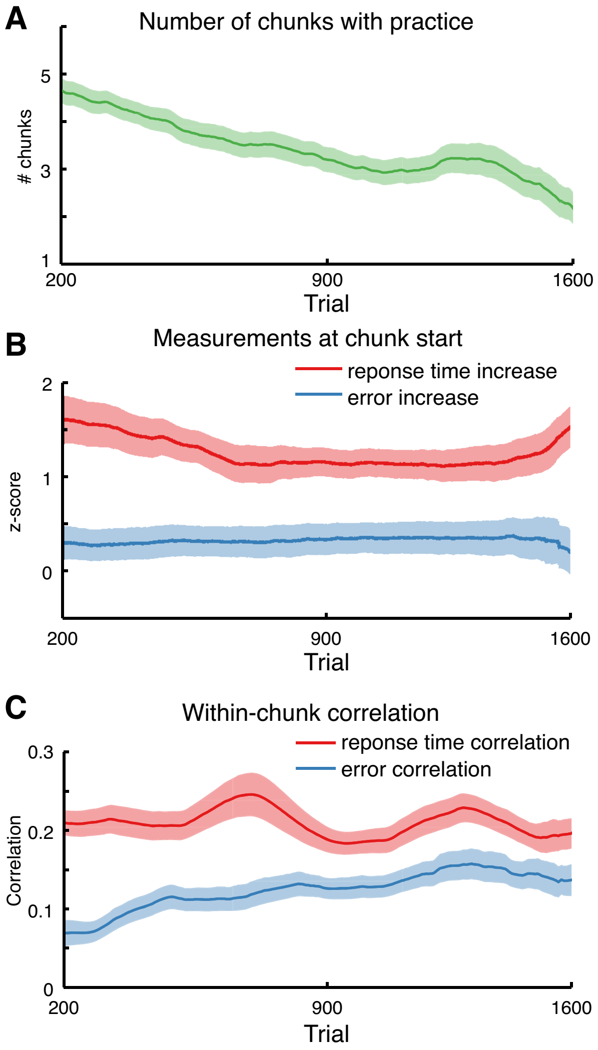

Fig. 5. Smoothed evolution of key features for DSP tasks on humans (±SE). *A*: mean number of chunks as a function of trial. Number of chunks used by subjects decreased over time *B*: standardized increase of response times and errors at the beginning of chunks. The reaction times signal is bigger than the error signal across the task. *C*: running mean correlation of response times and errors within chunks. Response times correlation stays strongly correlated within chunks while error correlations become a strong signal at the end of the task. The evolution of the error correlation within chunks is a novel feature revealed by the algorithm.

promises to help us better understand how motor memory is organized in sequence production tasks and how the central nervous system improves performance through practice.

Any model needs to make simplifying assumptions, and we want to discuss their effect. First, there could be different ways in which the central nervous system structures chunking. For example, we have assumed that, at any point during training, there is only one underlying chunking structure that we need to uncover. In principle, different brain regions could chunk differently, and chunking in one brain area could affect reaction times while the other affects error probabilities. However, our finding that response times can predict errors and vice versa suggests that at least some of the chunking structure is shared.

The data we analyzed come from experiments with several interleaved sequences, which is ignored by our analysis. We believe, however, that the interaction across sequences does not strongly affect our results. First, the parameters obtained during the detrending suggested that the effect of learning is essentially equal across subjects and sequences. Second, the sequences were originally created so that they do not have regularities (e.g., 121, 123, 11), lowering the potential for across-sequence interactions. Interference across sequences, however, could be potentially added by considering higher order hidden Markov models or hidden semi-Markov Models (Murphy and Paskin 2002). It seems unlikely that the structures we uncovered were produced by the interleaved sequences.

Previous studies have examined possible brain substrates of chunking behavior in the DSP task. Distinct areas have been found to correlate with chunk splitting (left-hemisphere front parietal network) and chunk concatenation (bilateral sensorimotor putamen) (Wymbs et al. 2012). Other studies have related the ability of humans to learn with the "flexibility" they have to change the representation of motor sequences (Bassett et al. 2011, 2013). Taken together, these results suggest that the sequence production tasks are a powerful, yet simple, paradigm to study learning and skill acquisition and its relationship to the brain. A tool like the one presented here can be used to pinpoint more precisely individual differences in learning performance as a result of improved chunking inference. A potential extension could penalize chunking structures that have either too few or too many chunks (Verwey 1994). This modeling assumption would add to the intuition that remembering many chunks adds to the cognitive load and therefore is less likely to occur (e.g., Verwey 1996), and also having one long chunk is unlikely because it would require the motor system to store the entire sequence in one chunk of memory (Sternberg et al. 1978). In our algorithm, this notion could be added by regularizing the initial distribution (*Eq. 3*) with an appropriate prior that penalizes too few or too many chunks.

We have introduced a coherent and statistically meaningful approach for dealing with chunking inference in sequence production tasks. While alternative methods can provide estimates of the chunking structure (e.g., Wymbs et al. 2012; Bassett et al. 2013; Mucha et al. 2010; Fortunato 2010), they start from a time-invariant evolution of chunking structure and do not deal with multiple features and correlations. Our algorithm produce estimates for *each trial*, by design, while other approaches provide an estimate for a window of trials whose size needs to be provided (Verwey 1996). Also, since our method is a fully generative model of how data are produced, it can gracefully deal with missing data, and it could be extended to provide estimates and predictions of any random variable with Bayesian credible intervals ("errors bars").

The fundamental ideas used in our algorithm also provide a framework to understand other motor learning tasks beyond DSP. For example, our algorithm is amenable to include hierarchical and non-Markovian probabilistic structures (e.g., see Heller et al. 2009). Our algorithm runs relatively fast (analyzing 2,000 trials requires around a minute on an ordinary desktop computer) and can be easily brought into the analysis of more complex tasks that involve varying, continuous motor command execution (e.g., grasping or reaching). Similar to the work that has been done in DSP tasks, our algorithm may serve to improve understanding of how motor memory is consolidated in general by using the same probabilistic principles we used here.

GRANTS

This work was supported by Public Health Service Program Project Grant NS-44393 (to K. Kording, P. L. Strick, R. S. Turner, and S. T. Grafton) and NS-074044 (to K. Kording).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: D.E.A. and K.K. conception and design of research; D.E.A. and K.K. analyzed data; D.E.A. and K.K. interpreted results of experiments; D.E.A. and K.K. prepared figures; D.E.A. and K.K. drafted manuscript; D.E.A., N.F.W., N.P., R.S.T., P.L.S., S.T.G., and K.K. edited and revised manuscript; D.E.A., R.S.T., P.L.S., S.T.G., and K.K. approved final version of manuscript; N.F.W., C.A.R., N.P., R.S.T., P.L.S., and S.T.G. performed experiments.

REFERENCES

Abrahamse EL, Ruitenberg MF, de Kleine E, Verwey WB. Control of automated behavior: insights from the discrete sequence production task. *Front Hum Neurosci* 7: 82, 2013.

Adams JA. Learning of movement sequences. *Psychol Bull* 96: 3, 1984.

Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST. Dynamic reconfiguration of human brain networks during learning. *Proc Natl Acad Sci USA* 108: 7641–7646, 2011.

Bassett DS, Wymbs NF, Rombach MP, Porter MA, Mucha PJ, Grafton ST. Task-based core-periphery organization of human brain dynamics. *PLoS Comput Biol* 9: e1003171, 2013.

Clerget E, Poncin W, Fadiga L, Olivier E. Role of broca's area in implicit motor skill learning: evidence from continuous theta-burst magnetic stimulation. *J Cogn Neurosci* 24: 80–92, 2012.

Crossman ER. A theory of the acquisition of speed-skill. *Ergonomics* 2: 153–166, 1959.

Desmurget M, Turner RS. Testing basal ganglia motor functions through reversible inactivations in the posterior internal globus pallidus. *J Neurophysiol* 99: 1057–1076, 2008.

Desmurget M, Turner RS. Motor sequences and the basal ganglia: kinematics, not habits. *J Neurosci* 30: 7685–7690, 2010.

Ericsson KA. The influence of experience and deliberate practice on the development of superior expert performance. In: *The Cambridge Handbook of Expertise and Expert Performance*, edited by Ericsson KA, Charness N, Feltovich P. Cambridge MA: Cambridge Univ. Press, 2006, p. 685–706.

Fortunato S. Community detection in graphs. *Phys Rep* 486: 75–174, 2010.

Gentner DR. Timing of skilled motor-performance–tests of the proportional duration model. *Psychol Rev* 94: 255–276, 1987.

Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. New York: Springer, 2009.

Heathcote A, Brown S, Mewhort D. The power law repealed: the case for an exponential law of practice. *Psychon Bull Rev* 7: 185–207, 2000.

Heller KA, Teh YW, Gorur D. Infinite hierarchical hidden Markov models. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Clearwater Beach, FL: *Journal of Machine Learning Research*, vol 5, 2009.

Jordan MI. Graphical models. *Stat Sci* 19: 140–155, 2004.

Logan GD, Bundesen C. Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *J Exp Psychol Hum Percept Perform* 29: 575, 2003.

Matsuzaka Y, Picard N, Strick P. Skill representation in the primary motor cortex after long-term practice. *J Neurophysiol* 97: 1819–1832, 2007.

Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 63: 81, 1956.

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP. Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328: 876–878, 2010.

Murphy KP. *Machine Learning :a Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

Murphy KP, Paskin MA. Linear-time inference in hierarchical HMMs. *Adv Neural Info Proc Syst 2*: 833–840, 2002.

Newell A, Rosenbloom PS. Mechanisms of skill acquisition and the law of practice. In: *Cognitive Skills and Their Acquisition*. Pittsburgh, PA: Carnegie Mellon Univ, 1981, p. 1–55.

Newell KM. Motor skill acquisition. *Annu Rev Psychol* 42: 213–237, 1991.

Parr RE. *Hierarchical Control and Learning for Markov Decision Processes* (PhD thesis). Berkeley, CA: Univ of California, 1998.

Pavlov I. *Conditioning Reflexes*. New York: Oxford Univ Press, 1927.

Skinner BF. *The Behavior of Organisms: an Experimental Analysis* (illustrated, reprint ed.). New York: D. Appleton-Century, 1938.

Sternberg S, Monsell S, Knoll RL, Wright CE. *Information Processing in Motor Control and Learning*. New York: Academic, 1978, p. 117–152.

Thorndike EL. Animal intelligence: an experimental study of the association processes in animals. *Psychol Rev Monograph* 2: 1–8, 1898.

Todorov E, Li W, Pan X. From task parameters to motor synergies: a hierarchical framework for approximately optimal control of redundant manipulators. *J Robot Syst* 22: 691–710, 2005.

Verwey WB. Evidence for the development of concurrent processing in a sequential keypressing task. *Acta Psychol* 85: 245–262, 1994.

Verwey WB. Buffer loading and chunking in sequential keypressing. *J Exp Psychol Hum Percept Perform* 22: 544, 1996.

Verwey WB, Dronkert Y. Practicing a structured continuous key-pressing task: motor chunking or rhythm consolidation? *J Mot Behav* 28: 71–79, 1996.

Welch LR. Hidden Markov models and the Baum-Welch algorithm. *IEEE Info Theory Soc Newsletter* 53: 1–14, 2003.

Wymbs N, Bassett D, Mucha P, Porter M, Grafton S. Differential recruitment of the sensorimotor putamen and frontoparietal cortex during motor chunking in humans. *Neuron* 74: 936–946, 2012.