E pluribus unum but how? Chunking as a rational solution to the speed-accuracy trade-off

Shuchen Wu[1,*], **Noémi Éltető**[2], **Ishita Dasgupta**[3], and **Eric Schulz**[1]

[1]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3]Google DeepMind, New York City, New York, USA
[*]shuchen.wu@tue.mpg.de

ABSTRACT

When exposed to perceptual and motor sequences, people are able to gradually identify patterns within and form a compact internal description of the sequence. One proposal of how sequences can be compressed is people's ability to form chunks. We study people's chunking behavior in a serial reaction time task. We relate chunk representation with sequence statistics and task demands, and propose a rational model of chunking that rearranges and concatenates its representation to jointly optimize for accuracy and speed. Our model predicts that participants should chunk more if chunks are indeed part of the generative model underlying a task and should, on average, learn longer chunks when optimizing for speed than optimizing for accuracy. We test these predictions in two experiments. In the first experiment, participants learn sequences with underlying chunks. In the second experiment, participants were instructed to act either as fast or as accurately as possible. The results of both experiments confirmed our model's predictions. Taken together, these results shed new light on the benefits of chunking and pave the way for future studies on step-wise representation learning in structured domains.

Introduction

William James famously said that we are born into a "blooming, buzzing confusion", and that we escape that confusion by gradually making sense of the series of events we perceive. How we perceive a sequence of perceptual stimuli, process them, and extract underlying structure, is a fundamental question of psychological investigations. One proposal of how the blooming, buzzing confusion of seemingly disparate sequential events can become one cognitive unit is chunking[1–4]. Upon exposure to sequential stimuli, humans and animals can identify repeated patterns and segment sequences into chunks of patterns[5]. To this end, separate sequential elements merge into one cognitive entity. This cognitive entity is then recalled and identified as a whole[6]: a phenomenon known as *chunking*[7,8].

Chunking is a phenomenon spanning across sequence learning, grammar learning, visual and working memory tasks, and function learning, among others[8–12]. The ability to discover statistical regularities in sequences, and to identify them as discrete, disparate units of chunks enables us to form a compact and compressed memory representation[13], readily transferable to novel domains[14], and enables us to progress from novices to experts[15,16]. As primitive building blocks of cognitive construction units, a complex and lengthy sequence reduces to several chunks. This property facilitates compositionality in learning[17], communication of structure[18], hierarchical planning[19], and the organization of actions[7]. In short, chunking is a critical and universal learning phenomenon.

We study the interaction of chunking with sequence statistics and propose another benefit of chunking in a sequential task: the ability to more easily predict future outcomes and thereby act faster. We study this phenomenon in a serial reaction time task (SRT), a classical paradigm to study motor sequence learning[12,20–22]. In SRTs, sequences of instruction cues appear consecutively on the screen, after which participants react by pressing the corresponding key that maps to the cue. If particular patterns, for example, ABC, keep repeating, then grouping repeated chunks as a unit facilitates the prediction of upcoming instruction sequences. The detection of a chunk's beginning, in this case, A, implies that the within-chunk items B and then C will follow. This anticipation of the following elements of a given chunk can allow participants to anticipate what is coming next and thereby react faster[12,14]. Chunking sequence elements, however, can also come at a cost when the sequence is probabilistic. By assuming deterministic transitions between the within-chunk items AB, participants might lose fine-grained statistical information about single-item instructions and thereby occasionally miss between-chunk transitions such as AC. This, in turn, can decrease their accuracy.

We propose a model that trades off between speed and accuracy when performing SRTs. Our model calculates the utility of acquired chunk representations as a weighted sum of how well they capture the statistical structure in the SRT (accuracy)

and whether they permit faster responses (speed). Our model then iteratively decides whether or not to chunk consecutive items. This model makes two distinct predictions. First, in environments where deterministic chunks exist, adding them to the representation is beneficial because they speed up reaction times without losing accuracy. Thus, people should chunk more in environments with more or longer chunks. We test and verify this prediction in our first experiment by training participants on sequences containing underlying chunks. The results of this experiment suggest that subjects adapt their chunking behavior to the underlying chunks in the sequence. A second prediction is that it can be rational to learn chunks in cases where the underlying environment is non-deterministic and does not contain any chunks. Since chunking frequently co-occurring events improves reaction time at the cost of overall accuracy, chunking can be a strategy to act faster. Thus, as the utility of speed increases (at the cost of accuracy), participants might also chunk consecutive elements more often and learn longer chunks. We test and verify this prediction in a second experiment by training subjects on sequences generated from a first-order Markovian transition matrix with "illusory" chunks while instructing one group to focus on speed and the other group to focus on accuracy. The results of this second experiment suggest that the group focusing on speed chunked more than the group focusing on accuracy. Our results shed new light on the benefits of chunking and pave the way for future studies on structural inference in statistical learning domains.

Serial Reaction Time Task

We study chunking in a serial reaction time task (SRT, see Fig 1b). Participants are instructed to press keys corresponding to a sequence of cues that appear on the screen. The instruction cross turns green after a correct keypress and red after an incorrect keypress. The subsequent trial starts after a 500ms response-to-stimulus interval. The task starts with two baseline blocks followed by six training blocks and ends with two test blocks. Each block consists of 100 trials. For both experiments, the same generative mechanism produces the baseline and the test blocks. To study whether participants' chunking behavior adapts to task demands in an SRT task, we manipulate various properties of the training blocks to examine how they affect behavior in the test block, using the baseline block as a comparison. The observed differences between the test and baseline blocks reflect the changes in representations elicited by the training blocks.

The instruction sequences in the baseline and test blocks across both experiments are generated from a non-deterministic, first-order Markovian transition matrix between the four instruction keys. In particular, out of all 16 transitions specified between the four keys, the transitions from A to B and C to D are highly probable (P = 0.9), and the transitions from B to C and from D to A are medium probable (P = 0.7) (see Fig 1c). In this way, participants often observe reoccurring sequence segments such as AB and CD and could possibly perceive them as "illusory" chunks, even though no such chunks truly exist in the generative model.

We manipulate the training block sequences across the two experiments. In Experiment 1, three groups of participants were trained on sequences containing no chunks (independent), chunk AB (size 2 chunk), or chunk ABC (size 3 chunk). In Experiment 2, the same "illusory" transition matrix generates the training block sequences but the instructions differ across the two experimental groups. One group is instructed to respond as accurately as possible, while the other is instructed to respond as fast as possible. In order to control for motor effects due to hand and finger dominance, the instructions "A", "B", "C", "D" are randomly mapped to the keys "D", "F", "J", "K" for individual participants. In the next section, we discuss the predictions of our rational model of chunking for the two different experiments and their conditions.

A rational model of chunking

In the SRT, single instructions z out of an instruction set Z are presented sequentially. We told participants to press the corresponding key as soon as a new instruction appears. The subsequent instruction shows up in a fixed interval after a participant's completion of the previous trial. The model learns a set of chunks $C = \{c_1, ..., c_n\}$ and uses the set to parse the sequence. It evaluates the probability $P(c)$ of parsing each chunk c and the conditional probability $P(c_j|c_i)$ that c_j follows c_i for every pair of chunks.

The set of chunks C is initialized as the set of available single instructions Z at the beginning of all simulations. The model updates this set by potentially concatenating existing pairs of chunks in C. Adding a chunk expands the parsing horizon as the rest of the within-chunk items are predicted to deterministically follow the initiation item of the chunk. Therefore, the subsequent within-chunk items are expected to be upcoming in the following trials. The model's accuracy might diminish if the subsequent instructions are inconsistent with the predicted within-chunk items. We relate subsequent item predictions to reaction times in the next section, and then explain the process by which a rational model updates chunks based on the trade-off between reaction times and accuracy.

Accounting for Reaction Times

We use a linear ballistic accumulator (LBA) model to simulate reaction times (RT). LBAs are a common class of multi-choice models[23, 24]. In the LBA, each choice corresponds to an evidence accumulator, translated to each four possible key-presses in
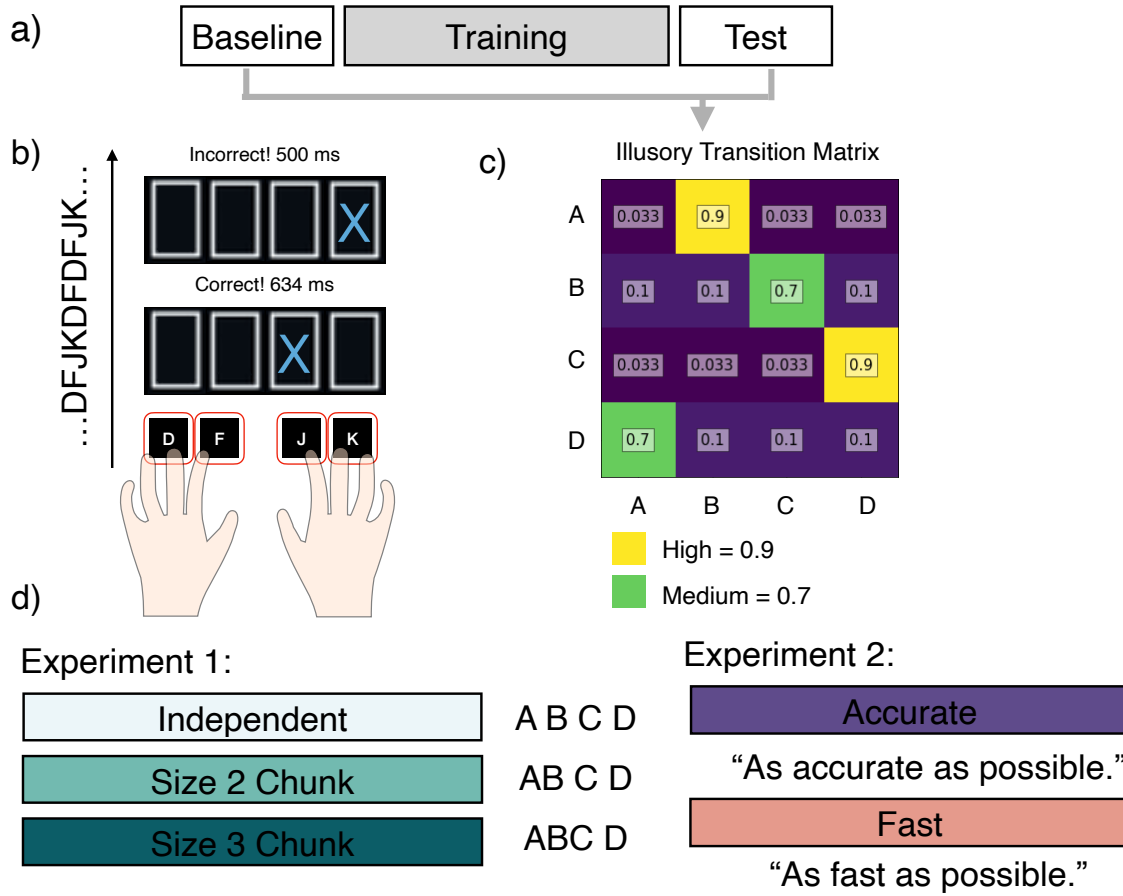
Figure 1. a) Task structure for both experiments. Six training blocks are sandwiched between two baseline and two test blocks. The baseline and test blocks contain instruction sequences generated from the "illusory" transition matrix in c). b) Participants are instructed to press the corresponding key on the keyboard according to trial-by-trial displayed instructions. They are given feedback on their performance, including accuracy and reaction times before the subsequent trial. c) A non-deterministic, "illusory" transition matrix of the four possible key-presses is used to generate instruction sequences for the baseline and test blocks for both experiments. The generative transition matrix with the two high (from A to B, C to D) and two medium transition probabilities (from B to C, D to A) produces "illusory" chunks that can be perceived as frequently occurring. To control the effect of habitual presses from consecutive fingers, a random mapping from "A", "B", "C", "D" to "D", "F", "J", "K", is generated independently for each participant. d) The instructions for training blocks differed between the two experiments and corresponding groups. In Experiment 1, participants were divided into three groups who learned independent, size 2, and size 3 chunks from a predefined set of chunks with equal probability. In Experiment 2, the instructions in the training group were generated from the "illusory" transition matrix. One group was instructed to act as accurately as possible and the other groups was instructed to act as fast as possible.

our task. At every trial of the SRT task, each evidence accumulator starts with an initial evidence $k = \log(P(z_i))$, which reflects the model's prediction on the upcoming instructions. The trials are divided into within-chunk trials and between-chunk trials. For a within-chunk trial, the prediction for the within-chunk item is the initial evidence for the accumulator $\log(1)$, the rest being $\log(\varepsilon)$. Note that the model still integrates information from the SRT instructions but with a high offset which biases it to choose the response which is consistent with the chunk, even if it is inconsistent with the instructed item. This term encourages the model to create longer chunks to reduce the average reaction time.

For a between-chunk trial, the initial evidence for each accumulator z_i is determined by the transition probability $P(c_i|c_j)$ of the chunk c_i that initiates with the accumulator z_i, given the previously parsed chunk c_j. All response accumulators start from the initial evidence, and drift towards the decision threshold with positive drift rates v_A, v_B, v_C, v_D sampled from a normal distribution with mean $v_{instruction}$ and standard deviation σ. To simulate the RT of a particular trial, the current instruction
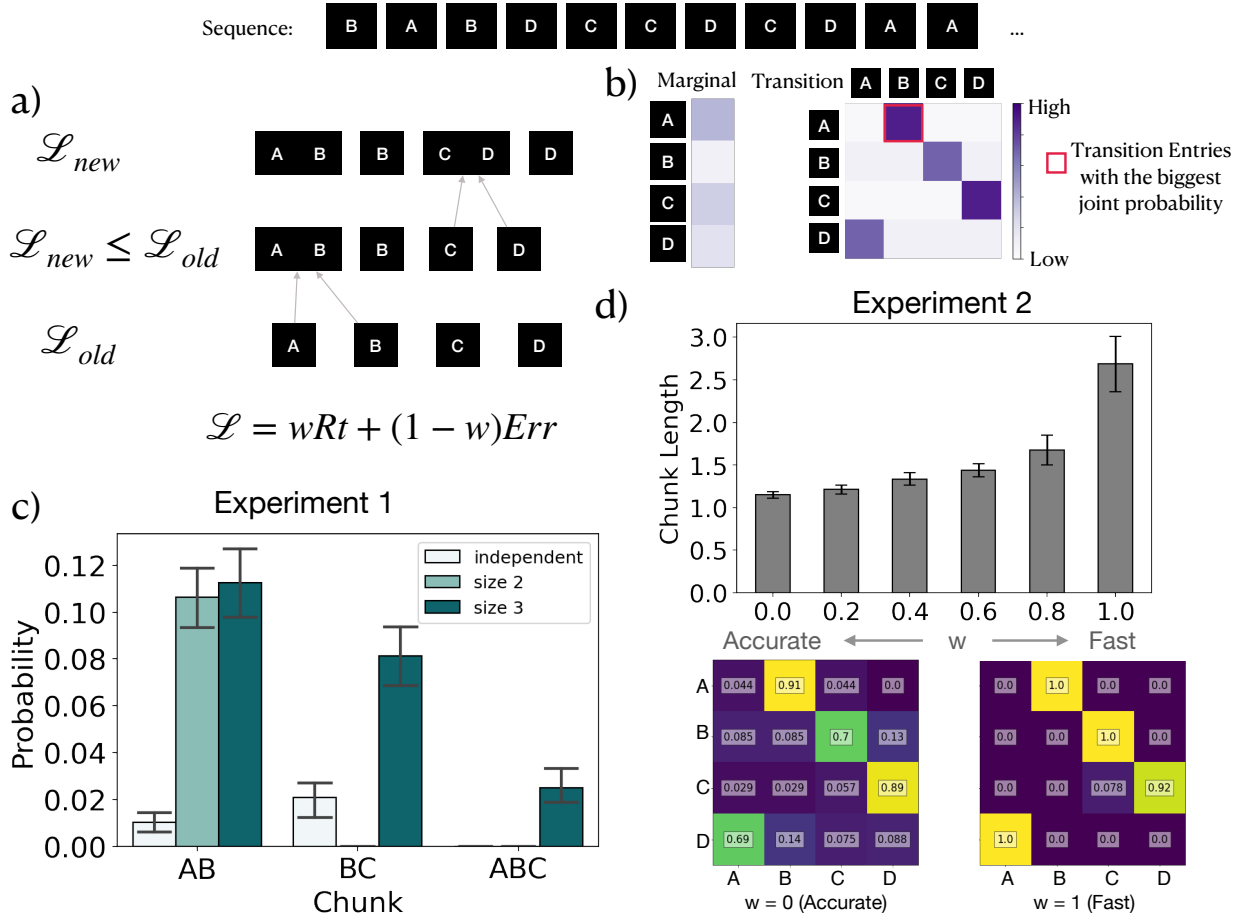
Figure 2. a) Chunking mechanism of rational model. The model keeps track of marginal and transitional probabilities among every pair of pre-existing chunks, and combines chunk pairs that yield the greatest joint probability as the next candidate to be chunked together. At the start, the four different keys are initialized to be the primitive chunks. A loss function that trades off reaction times and accuracy is evaluated on the pre-existing set of chunks. If a chunk update reduces the loss function, then the two pre-existing chunks are combined together. A parameter w determines how much more the model weighs a decrease of reaction times compared to an increase in accuracy. b) Example model simulations of learning instruction sequences of Experiment 1. Because the transition AB occurred frequently, the model proposes this transition as a possible chunk. c) Model simulation for Experiment 1. Bars represent the probability of a particular chunk parsed in a simulation over the whole experiment. Note that these bars can be arbitrarily increased by changing w while the qualitative results remain the same. d) Model simulation for Experiment 2. Top: Average chunk length of different simulations when increasing w from 0 (optimizing only accuracy) to 1 (optimizing only speed). As w increases the average chunk length increases, indicating that the model learns longer chunks when asked to care more about acting fast. Bottom: Transition probabilities learned by model with $w = 0$ and $w = 1$, corresponding to the rational maximization of accuracy and speed. If the model tries to act as accurately as possible, then it recovers the true transition probabilities of the "illusory" transition matrix. If the model tries to act as fast as possible, then sets the medium and high transition probabilities to be 1, i.e. deterministic. All results are averaged across 120 independent simulations. Error bars represent the standard error of the mean.

carries the highest drift rate $v_{instruction} = 0.5$. The evidence accumulators corresponding to the other instructions have an equal but lower drift rate $v_{\neg instruction} = \frac{1 - v_{instruction}}{3}$. The drift rates for all accumulators sum up to 1. For example, if the current instruction is A, then $v_A = 0.5$, $v_B = v_C = v_D = \frac{0.5}{3}$. Evidence accumulation terminates when a positive response threshold b is first crossed by any accumulator. The accumulator that crosses the decision threshold first becomes the overt response, and the time it takes to reach the decision threshold is the simulated RT on that trial. In all of the model simulations, we use the same $v_{\neg instruction} = 0.5$, decision threshold $b = 1$, $\varepsilon = 0.01$, and standard deviation $\sigma = 0.6$ across all accumulators.

Balancing Speed and Accuracy

We assume that chunking enables participants to predict upcoming instructions further into the future and thereby to react faster by initializing their evidence at a higher starting point. However, chunking also bears a risk of making mistakes when the upcoming instructions are not the subsequent items within a chunk. We formulate this speed-accuracy trade-off using the loss function

$$\mathscr{L} = wRt + (1-w)Err, \tag{1}$$

where Rt is the average reaction time in the SRT, given a learned chunk representation and sequence, and Err is the average error rate. w is a parameter that specifies the trade-off between accuracy and reaction time. When $w = 0$, only the reaction time term Rt occupies the loss, and when $w = 1$, the error term Err dominates.

Based on the LBA reaction time simulation, the average reaction time of parsing chunk c_j after previously having parsed the chunk c_i is $\frac{rt_{between}(c_j,c_i)+(|c_j|-1)rt_{within}}{|c_j|}$. $|c_j|$ is the length of the chunk. The reaction time on the first item is denoted as $rt_{between}(c_j,c_i)$, since $P(c_j|c_i)$ influences the evidence accumulation for this between-chunk key press and only the boundary of the chunk contains transition uncertainty and contributes to the slow down of reaction times. As the initial chunk item determines the chunk identification, the subsequent reaction time to press within-chunk keys in c_i is denoted as rt_{within}. This term does not depend on c_i as the procession to the within-chunk items contains no uncertainty. Taken together, the average reaction time can be formulated as follow, averaging the probability of parsing each acquired chunk c_j given the previously parsed chunk c_i

$$Rt = \sum_{c_i \in C} P_C(c_i) \sum_{c_j \in C} P_C(c_j|c_i) \left[\frac{rt_{between}(c_j,c_i)+(|c_j|-1)rt_{within}}{|c_j|} \right], \tag{2}$$

Similarly, if we formulate $R_{LBA}(z_j)$ as the response choice of the LBA model when the instruction is z_j, then we can denote an error occurrence as $\mathbb{1}[z_j \neq response(z_j)]$, which is an indicator function that becomes 1 when the instruction z_j is inconsistent with the LBA response. The average error rate can be evaluated by averaging the error rate with the probability of single-element transitions from the generative model P_I, enabling the formulation of the expected error rate as

$$Err = \sum_{z_i \in \{A,B,C,D\}} P_I(z_i) \sum_{z_j \in \{A,B,C,D\}} P_I(z_j|z_i) \mathbb{1}[z_j \neq R_{LBA}(z_j)] \tag{3}$$

This utility function, therefore, induces a trade-off between being accurate (predicting elements correctly) and being fast (finding a chunk representation to predict further ahead and speed up one's reaction time). Together, these parts of the loss are used to evaluate the utility of a chunk representation under specific task demands.

The Rational Update of Chunking

The model updates the chunk representation rationally by concatenating chunks within the chunk set C that induce a lower loss. C is initialized with single sequential items $\{A,B,C,D\}$. For one set of chunks C, the model evaluates the marginal probabilities of each chunk $P_C(c_i), c_i \in C$ and the transition probability $P_C(c_j|c_i)$ of parsing chunk c_j after having parsed chunk c_i. $P_C(c_i)$ and $P_C(c_j|c_i)$ are stored in the marginal and transition probability matrix as shown in Figure 2a. The marginal and transition probability is evaluated empirically over an entire sequence parse using chunks in C.

We can calculate the joint occurrence probability of concatenating chunk c_i with c_j as $P(c_i,c_j) = P_C(c_j|c_i)P_C(c_i)$. The chunk pair c_i, c_j with the highest joint probability is suggested as a new chunk to replace c_i to form a new to the set of chunks C_{new}. As the initiation of c_i is predictive of the subsequent chunk items. For example, an addition to $\{A,B,C,D\}$ could be a new chunk AB. The new chunk AB then replaces A and the new proposed set of chunks C_{new} becomes $\{AB,B,C,D\}$.

We then compute whether C_{new} is accepted to replace the original set of chunks C. The acceptance depends on whether the new set of chunks C_{new} and the induced reaction time in addition to the marginal and transition probabilities upon parsing the sequence leads to a lower loss. In case it is so, C_{new} replaces C, which becomes the basis of proposing the next chunk. This chunk proposal process continues until a fixed iteration number.

Model Predictions for Experiment 1

We first examined the model's chunk learning behavior on the three groups of Experiment 1. In this simulation, the underlying generative model either contained no chunks (independent), the chunk AB (size 2 chunks), or the chunk ABC (size 3 chunks). We then fixed the trade-off parameter w to optimize accuracy more than speed by setting it to $w = 0.2$. Figure 2b shows the probability of chunk AB, BC, and ABC being learned as subchunks by the rational model of chunking over 120 simulations in total. The model uses the entire sequence to learn its chunk representation in each simulation. With the same trade-off between

speed and accuracy, the rational chunking model trained on sequences with size 2 and 3 chunks has a higher probability of learning AB as a subchunk than a model trained on the independent sequence. Chunk BC has a higher probability of being learned by the model trained on sequences containing size 3 chunks than models trained on sequences with independent instructions or sequences with size 2 chunks. Only the model trained on sequences containing size 3 chunks learned about the chunk ABC. Taken together, these simulations predict that participants in the different conditions will be more likely to learn the corresponding chunks than participants for whom a chunk is not part of the training sequence.

Model Predictions for Experiment 2

We examined the model's chunk learning behavior for Experiment 2. According to our model, changing the trade-off between accuracy and speed translates to changing the cost function's w away from 0 and towards 1. We therefore simulated the behavior of our model with changing w (Fig 2b). As w goes from 0 to 1, i.e. the cost function shifts from minimizing the model's error rate to minimizing its reaction time, the average length of chunks learned by the model increases. Thus, our model predicts that participants in the fast group, which demands speedier responses, should learn longer chunks as compared to participants in the accurate group. Evaluating the single-element transition probability with $w = 0$ and 1 (Fig. 2a) shows that if only accuracy is the optimization goal of the cost function, then the model preserves the original transition matrix. However, if the model optimizes for speed, then it learns a polarized transition probability where all the high and medium single element transitions attract more probability mass, i.e. are closer to 1. Correspondingly, the remaining probabilities are closer to 0. Thus, as the high and medium transitions are more integrated into the chunks, this gives the model a speed-up in its reaction times, because it can start its evidence accumulation at a higher initial point. This comes at the cost of accuracy, because the initialization may be incorrect.

Experiment 1: Learning about true chunks

In Experiment 1, we test the model's prediction that chunking behavior adapts to the statistics of the sequence. When chunks are used to generate the sequence, participants should learn more than those trained on sequences without chunks.

We randomly assigned 122 participants to one of three groups. Training blocks differed amongst the three groups, while the baseline and the test blocks remained the same. For the training blocks, the independent group practiced sequences that contained no chunks, the size 2 group practiced sequences with chunk AB, and the size 3 group practiced sequences with chunk ABC, as shown in Figure 1d. The baseline and test blocks were sequences generated by the illusory transition matrix in Figure 1c. The main prediction was that if people have learned chunks present in the training blocks, then they will use them even in the test blocks. We measured this by examining differences in accuracy and reaction times from the baseline block. We also used Experiment 1 to validate several of our empirical measures of chunking which we will use in Experiment 2.

Manipulation check

We first checked if participants' behavior during the training blocks reflected the underlying chunks in the generative model. In particular, we tested whether the size 2 group showed evidence for learning chunk AB, and the size 3 group learning chunk ABC. We used a Gaussian mixture model to categorize reaction times of each response from the same participant into fast "within" or slow "between" chunk transitions, based on the assumption of a within-chunk speed-up. This method gave us a glimpse into how the action sequence was partitioned by participants, reflecting their internal representation of chunks (more details in Methods). We then counted the number of times chunks AB and ABC showed up in the training block, denoted as N_{AB} and N_{ABC}. If the size 2 group and the size 3 group had learned chunk AB and ABC, separately, then N_{AB} should be higher for these two groups than the independent group, and N_{ABC} should be higher for size 3 group than the other two. Figure 3a shows the average N_{AB} and N_{ABC} returned by this analysis across the three conditions during the training blocks.

For N_{AB}, fitting a linear regression model using condition as the independent variable and the number of chunks N_{AB} as the dependent variable showed a significant effect of condition ($F(2) = 45.02$, $p < 0.001$). N_{AB} was higher for both the size 2 group ($\hat{\beta} = 59.43$, $t(139) = 8.20$, $p < 0.001$) and the size 3 group ($\hat{\beta} = 59.39$, $t(139) = 8.32$, $p < 0.001$) than for the independent group. This means that training on instruction sequences that contained either AB or ABC chunks induced participants to learn AB as a chunk.

To investigate differences in the acquisition of the ABC chunk between groups, we repeated the same regression with N_{ABC} as the dependent variable. We found a significant effect of groups ($F(2) = 71.45$, $p < 0.001$), indicating that participants' responses reflected ABC chunks more often in both the size 2 ($\hat{\beta} = 9.40$, $t(139) = 1.89$, $p = 0.06$) and size 3 ($\hat{\beta} = 54.17$, $t(139) = 11.07$, $p < 0.001$) than in the independent group. Interestingly, we observed higher N_{ABC} with the size 2 group than in independent group. This can be because building on top of a previously learned chunk (AB $\rightarrow$ ABC) is more accessible for the size 2 group than the independent group (as the independent group needs to learn chunk AB first, then ABC). Furthermore, N_{ABC} was significantly higher for the size 3 than the size 2 group ($\hat{\beta} = 44.76$, $t(139) = 9.25$, $p < 0.001$), suggesting that
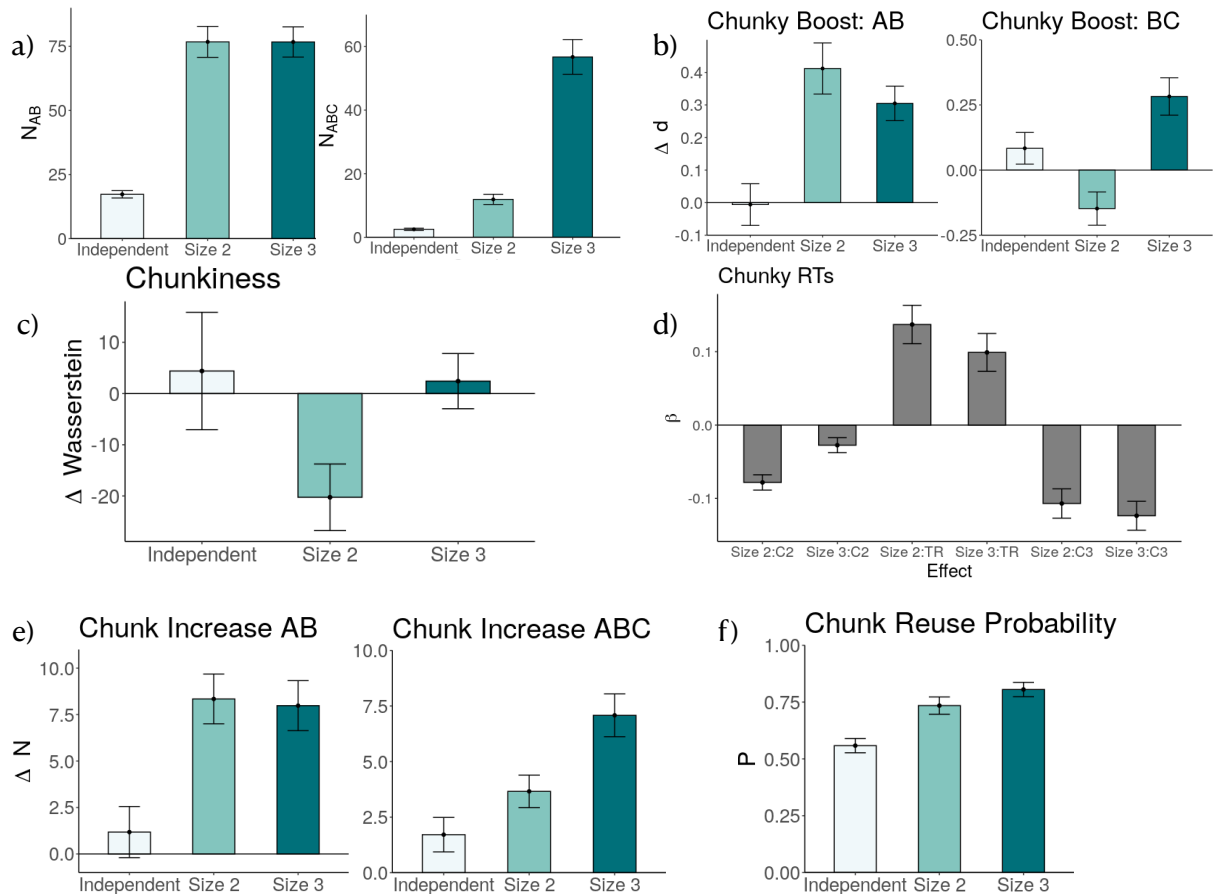
Figure 3. Results of Experiment 1. a) Manipulation check. The number of chunks AB and ABC learned by participants during the training blocks by group. Chunks were retrieved using a categorization of between- and within-chunk transitions by a mixture of Gaussians analysis of participants' reaction times. b) Chunky Boost of size 2 chunks AB and BC by group. A chunky boost is measured by the relative change of Cohen's *d* between baseline and test blocks for the highly and medium probable transitions. c) Chunkiness of size-3 chunks ABC. Chunkiness is measured by the relative change of Wasserstein distance between the baseline and test blocks of between-chunk reaction times of all possible size 3 chunks. d) Regression coefficients of interaction effects between condition and size 2, size 3, and true transition probabilities on reaction times during the final test blocks. e) Chunk increase from the baseline to the test blocks by group for chunk AB and chunk ABC. Chunk increase is measured by the number of returned chunks from the mixture of Gaussians analysis. f) Chunk reuse probability by group. Chunk reuse probability was calculated based on whether or not part of an earlier chunk were used in a later chunk that occurred within the next 30 trials. For all plots, error bars indicate the standard error of the mean.

training on instruction sequences that contained ABC chunks resulted in the strongest tendency of participants to learn ABC as a chunk.

Given these results, we conclude that our experimental manipulation of the three groups induced the intended behavior during the training blocks.

Chunky Boost

To study how training with underlying chunks affected the learning of chunks AB and BC, we looked at the group-wise reaction time change from the baseline blocks to the test blocks.

The reaction time of within-chunk items is typically faster than between-chunk items[8, 12, 25]. Based on this established result, we formulated a measure of chunky boost (Fig. 3b). This measure computes the speed-up of reaction times from the baseline to the test blocks when comparing within-chunk items with controlled between-chunk items.

We calculated a signed effect size of Cohen's d to evaluate the speed-up of within-chunk reaction times from the baseline to the test blocks, i.e., B within AB and C within BC. Cohen's d is a standardized measure of how far the means of two probability distributions are apart. In this case, these two distributions are the reaction time in the baseline blocks $rt_{baseline}$ and the reaction time in the test blocks rt_{test}. We used a signed version of Cohen's d to convey the relative change of the reaction time distributions. In the case of speed-ups, Cohen's d will be positive; the case of slowing-downs, it will be negative. For the controlled between-chunk items, we evaluated the signed Cohen's d on AA, AB, AC for chunk AB; and on BA, BB, BD for chunk BC. Finally, we arrived at the chunky boost measure Δd by subtracting the relative speed-up of AB and BC from their corresponding control chunks between the baseline and the test blocks. Detailed equations on this calculation can be found in the Supporting Information. Figure 3b shows the Chunky Boost of AB and BC across the three groups.

For chunk AB, fitting a linear model onto participants signed Cohen's d change showed a significant effect of group ($F(2) = 10.613, p < 0.001$); participants in the size 2 group had a higher relative change of Cohen's d than the independent group ($\hat{\beta} = 0.41, t = 4.41, p < 0.001$). Thus, training on the chunks with size 2 made the size 2 group respond to B faster after having seen item A. Additionally, participants in the size 3 group also had a higher relative change of reaction times responding to chunk AB than the independent group ($\hat{\beta} = 0.31, t = 3.35, p = 0.001$), showing that their reaction to B also sped up.

For chunk BC, fitting a linear model onto Δd with group as the independent variable also showed a significant effect ($F(2) = 10.802, p < 0.001$). Interestingly, the size 2 group had a negative chunky boost to BC ($\hat{\beta} = -0.23, t = -2.44, p = 0.02$), showing a relative reaction time slow-down compared to the control. This effect was expected because identifying B as the end of a chunk and the transition to C, therefore, being a "between-chunk" transition: a slow down in between-chunk reaction times was also observed in previous SRT experiments[26]. Relative to the independent group, the size 3 group had a significantly positive Cohen's d change ($\hat{\beta} = 0.20, t = 2.14, p = 0.03$). This shows that learning chunks changes the reaction time profile of this group, their response to C upon previous instruction B was speeding up their reaction times.

In summary, participants' reaction times changed in a predicable fashion, with the independent group not getting faster for either AB or BC, the size 2 group becoming faster for AB and slower for BC, and the size 3 group becoming faster for both AB and BC. These observations confirmed previous work studying chunking in SRT tasks, which has argued that RTs in structured sequences decreases more quickly than in non-structured sequences[12].

Chunkiness

We formulated a measure of chunkiness to evaluate the homogeneity of within-chunk reaction times for the chunk ABC (Fig. 3c). If participants have learned the chunk ABC, then the distributions of within-chunk reaction times should become more similar to each other. We used a relative change in Wasserstein distances to measure the homogeneity between the reaction times of pressing B and C after following item A.

The Wasserstein distance is also known as the "earth mover's" distance. It can be seen as the minimum amount of "work" required to transform one distribution into another. "Work" is the amount of distribution weight that must be moved multiplied by the distance (see also Supporting Information).

We measured the Wasserstein distance between the reaction time distribution rt_B and rt_C following item A. If participants have learned ABC as a chunk during the training blocks, then the two distributions should become more homogeneous in the test blocks as compared to the baseline blocks. We subtracted the Wasserstein distance evaluated on the test blocks from that evaluated on the baseline blocks to calculate this change of reaction time homogeneity. This relative change of Wasserstein should be positive if reaction times became more homogeneous in the test blocks. To control for an overall increase of homogeneity as the task progressed, we compared this change of Wasserstein with sub-sequences that were not ABC. The resulting measure of "chunkiness" can be seen as the relative change of the Wasserstein distance of chunk ABC compared to control transitions.

Chunkiness differed significantly between the three conditions ($F(2) = 3.20, p = .04$). In particular, the size 2 group had a negative relative Wasserstein shift ($\hat{\beta} = -26.92, t(137) = -2.37, p = 0.02$). This means that the size 2 group's reaction time

distribution became less homogeneous after training, indicating that the reaction time to press B deviated more from C. This was expected as for the size 2 group, pressing B and C after A should be one within and one between-chunk reaction time. On the other hand, the change of Wasserstein distance between the size 3 and the independent group condition was not significant, even though we would have expected this group to become more homogeneous in their reaction times. One reason for this surprising result could be that, while the reaction distribution upon the instructions "B" and "C" became closer to each other relative to control, the shift may have not uniformly impacted the calculation of Wasserstein distance.

In summary, we verified the prediction that the size 2 group had less homogeneous transition times within the chunk ABC than the other groups. However, we did not observe an increased chunkiness for the size 3 group, possibly due to non-uniform speed-ups of RTs.

Reaction Time Regression

We also studied the interaction between transition probabilities and group influencing participants' reaction times during the test blocks (Fig. 3d), when all groups had to go through sequences containing the illusory chunks. We fitted a linear mixed-effects regression using log-reaction times as the dependent variable, assuming a random intercept for each participant. The independent variables were the true (TR), size 2 (C2), and size 3 (C3) transition probabilities, group, as well as interaction effects between group and the transition probabilities. The size 2 (C2) and size 3 (C3) transitions reflected the deterministic transitions elicited by chunk formations.

The best regression contained the predicted transition probabilities as well as three interaction effects with groups ($\chi^2(8) = 129.6$, $p < 0.001$). The first interaction was between TR and the size 2 group ($\hat{\beta} = 0.13$, $t(2504) = 5.24$, $p < 0.001$), showing that the effect of TR learned by the size 2 group was up-weighted. The interaction between TR and the size 3 group was also significantly up-weighted ($\hat{\beta} = 0.10$, $t(2504) = 3.84$, $p = 0.39$). TR transition probabilities slowed down the reaction times for the size 2 and size 3 groups more than for the independent group.

The interaction was down-weighted between the C2 chunky transition probabilities and the size 2 group ($\hat{\beta} = -0.08$, $t(2504) = -7.51$, $p < .001$) and the size 3 group ($\hat{\beta} = -0.03$, $t(2504) = -2.68$, $p = 0.007$). This indicates that C2 transition probabilities sped up the reaction times for the size 2 and the size 3 group more than for the independent group.

Finally, the interaction was down-weighted between the C3 transition probabilities and the size 3 group ($\hat{\beta} = -0.12$, $t(2504) = -6.27$, $p < .001$), as well as an interaction between the C3 transitions and the size 2 group ($\hat{\beta} = -0.11$, $t(4967) = -5.34$, $p < .004$). These significant interaction effects indicate that C3 transition probabilities sped up the reaction times for the size 2 and the size 3 group more than for the independent group.

In summary, we found predictable relations between participants' reaction times and the transition probabilities implied by the chunks across groups. In particular, C2 transitions were significantly related to speed-ups for the size 2 and the size 3 groups, while C3 transitions were only related to speed-ups for the size 3 group.

Chunk Increase

As another method to study how the training blocks affected chunking behavior, we measured the number of times each participant chunked AB and ABC from their chunk profile computed by the mixture of Gaussians classification method during the baseline and the test blocks. We evaluated the increase ΔN of chunking AB and ABC from the baseline to the test blocks. Figure 3e shows ΔN_{AB} and ΔN_{ABC} across the three groups.

Fitting a linear model setting ΔN of AB as the dependent variable, and group as an independent variable showed a significant effect of group ($F(2) = 8.64$, $p < 0.001$). Compared to the independent group, the size 2 group ($\hat{\beta} = 7.16$, $t(139) = 3.68$, $p < 0.0.001$) and the size 3 group ($\hat{\beta} = 6.80$, $t(139) = 3.55$, $p < 0.0.001$) used the chunk AB significantly more in the test compared to the baseline blocks.

The same analysis for ΔN of chunk ABC also showed a significant effect of group ($F(2) = 10.63$, $p < 0.001$). The size 3 group used significantly more chunks ABC than the independent group ($\hat{\beta} = 5.36$, $t(139) = 4.53$, $p < 0.001$). Participants in the size 2 group also used more chunks ABC than the participants in the independent group ($\hat{\beta} = 1.94$, $t(139) = 1.62$, $p = 0.10$). Compared with the size 2 group, the size 3 group used ABC significantly more often ($\hat{\beta} = 3.42$, $t(139) = 2.92$, $p = 0.004$).

This analysis confirmed that training on instruction sequences with chunks increased participants' tendency to use those chunks in the test blocks that contained only illusory chunks.

Chunk Reuse

Finally, based on the individual chunking profiles identified by the mixture of Gaussians classification, we examined participants' chunk reuse probability across the three groups during the training blocks (Fig 3f). As the rational chunking model reuses previously learned chunks to construct new ones, we wanted to study whether participants' chunking behavior reflected this feature of our model.

The chunk reuse probability was evaluated on chunks of size three or bigger (excluding chunks with single item repetitions). Every time such a chunk occurred, we looked up 30 previous chunks backwards in a participants' learning history to check whether the current chunk was a sub-chunk of any previous chunks. Doing so, we arrived at a chunk reuse probability for each participant. Fig 3f shows the average chunk reuse probability across the three groups.

We found that the chunk reuse probability differed significantly between the groups. Fitting a linear model taking the reuse probability as the dependent variable and group as the independent variable showed a significant effect of condition ($F(2) = 13.99$, $p < 0.001$). Both the size 2 ($\hat{\beta} = 0.17$, $t(139) = 3.63$, $p < 0.001$) and the size 3 group ($\hat{\beta} = 0.24$, $t(139) = 5.17$, $p < 0.001$) group reused chunks significantly more often than the independent group. There was no significant difference between the size 2 and the size 3 group ($\hat{\beta} = 0.07$, $t(139) = 1.51$, $p = 0.13$).

The tendency to reuse previously learned chunks is consistent with how our model creates chunks, i.e. creating a new chunk by combining previously learned chunks. Apart from that, sequence statistics modulated participants' tendency to reuse chunks. Specifically, when the sequence contained embedded chunks that render reuse as beneficial to performance, participants tended to reuse previously acquired chunks more often than when the sequence only contained independent item instantiations.

The observation that participants reused previously acquired chunks echoes previous findings in the literature of transferring motor skills, which showed that people transfer chunks from a practiced sequence to a test sequence when there are shared chunks between the two[14]. In the current task, the reuse and transfer process was an ongoing learning behavior while participants practiced the training sequence.

Experiment 2: Learning chunks of different sizes to balance the speed-accuracy trade-off

In Experiment 2, we test the model prediction that chunk learning adapts to task demands. Participants should chunk more under time pressure, even given a sequence without chunks within.

We randomly assigned participants to one of two groups: the fast group and the accurate group. Both groups were trained on sequences generated from the "illusory" matrix that contained no true chunks but high and medium single item transitions (Figure 1c). The experiment structure was identical to the structure of Experiment 1: the training blocks were sandwiched between baseline and test blocks. The difference to Experiment 1 was that both groups received the same instruction to act "as fast and accurately" as possible during the baseline and the test blocks. Instruction and reward differed during the training blocks, during which the fast group was instructed to act as fast as possible and the accurate group to act as accurately as possible.

Manipulation check

We first assessed whether the instructions to be fast or accurate influenced participants' reaction times and accuracy during the training blocks. Shown in Figure 4a are the average reaction time and accuracy for the two groups. Fitting a linear mixed-effects regression onto participants' reaction times assuming a random intercept over individual participants showed a significant effect of group ($\chi^2(1) = 9.84$, $p = .002$), showing that participants in the fast group responded faster during the training blocks than participants in the accurate group ($\hat{\beta} = 81.71$, $t(113.93) = 3.19$, $p = .0001$). We also fitted a mixed-effects logistic regression of group to test whether participants responded correctly on each trial, adding a random intercept for each participant. This analysis also showed a significant effect of group ($\chi^2(1) = 9.67$, $p = .002$), with participants in the accurate group responding on average more accurately during the training blocks than participants in the fast group ($\hat{\beta} = 0.54$, $z = 3.18$, $p = .001$). Thus, we conclude that our experimental manipulation induced the intended behavior for the two groups during the training blocks.

Chunky boost

To study how the training blocks influenced the two groups to learn size two chunks from baseline blocks to the test blocks, we measured chunky boosts on the relative speed-up of reaction times responding to within-chunk items relative to between-chunk items.

Here, chunky boost was evaluated on the most frequently occurring size-2 chunks in the instruction sequence produced by the "illusory" transition matrix, i.e., AB, BC, CD, and DA, with AB, CD being size 2 chunks with high transition probability ($p = 0.9$), and BC and DA being those with medium transition probability ($p = 0.7$). The corresponding control chunks were size two subsequences that did not begin with the first chunk items. For example, the control chunks for AB were BB, CB, and DB. We conjectured that the fast group would learn more size two chunks with high and medium probability.

The chunky boost measured by a change of Cohen's d Δd is shown in Figure 4a. Fitting a linear mixed-effects regression onto participants' change of Cohen's d, assuming a random intercept over participants showed a significant effect of group ($\chi^2(1) = 7.25$, $p = .007$). Participants in the fast condition showed a greater relative boost in reaction times to chunky transitions as compared to participants in the accurate group ($\hat{\beta} = 0.24$, $t(73) = 2.71$, $p = .008$). We, therefore, concluded that participants in the fast group chunked more size two chunks than participants in the accurate group, as predicted by our model.
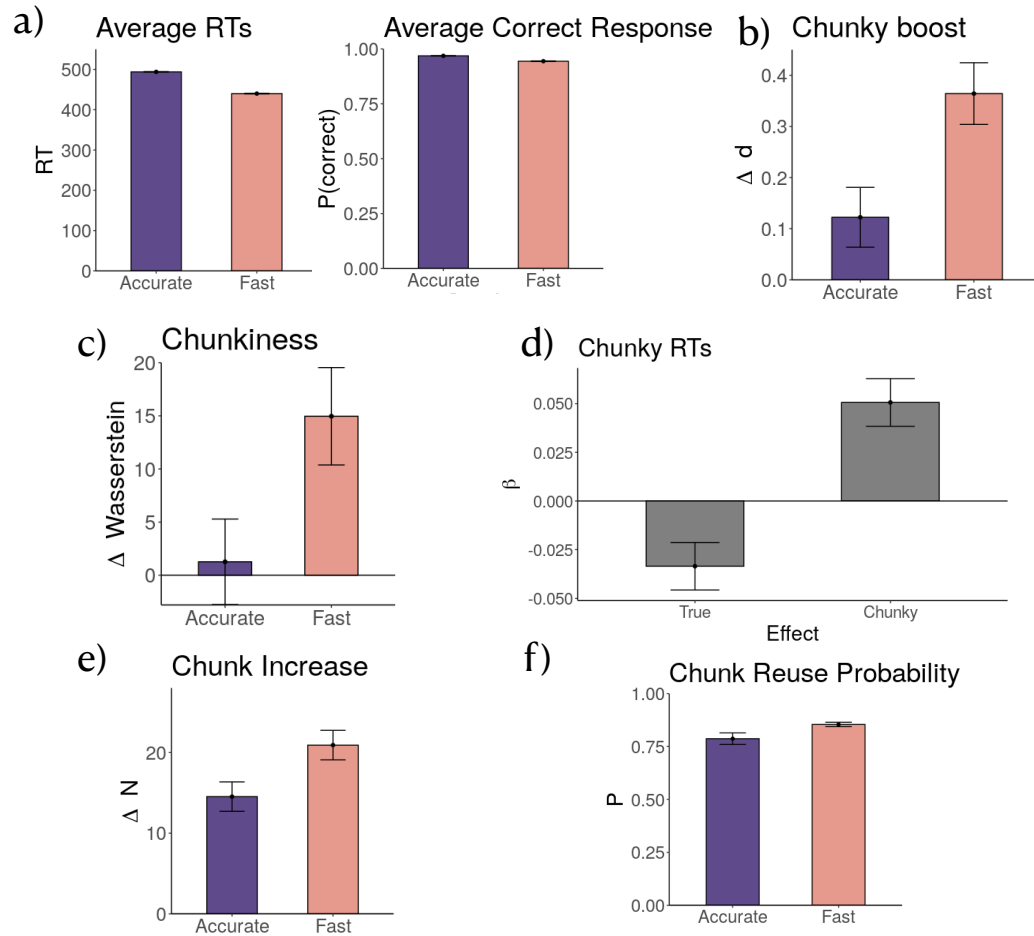
Figure 4. Results of Experiment 2. a) Manipulation check. Average reaction times and average response accuracy during training blocks by group. b) Chunky Boost of size 2 chunks as measured by change of Cohen's *d* by group evaluated on baseline and test blocks. The size 2 chunks include AB, BC, CD, and DA. c) Chunkiness measured by a relative change of Wasserstein distance of size 3 chunks including ABC, BCD, CDA, DAB between the baseline and the test blocks. d) Coefficient of interaction effect between chunky and true transition probabilities on reaction times during the test blocks. e) Chunk increase from the baseline to the test blocks by condition for size-2 (AB, CD, BC, DA) and for size-3 chunks (ABC, BCD, CDA, DAB). f) Chunk reuse probability by group. For all plots, error bars indicate the standard error of the mean.

Chunkiness

We also measured the reaction time homogeneity between the within chunk items of size three chunks using the same measure of chunkiness as in Experiment 1. This time, we examined this chunkiness measure on the most frequent size three chunks, including ABC, BCD, CDA, DAB as generated by the underlying "illusory" transition matrix. The Wasserstein distance was used to measure the shift of reaction time distributions between the 2nd within-chunk and the 3rd within-chunk items separately in the baseline and test blocks. The Wasserstein distance evaluated from the test block for size three chunks was subtracted from that evaluated on the baseline blocks. This time, the controls were the non-chunk subsequences of size three. The change in Wasserstein distance of the control chunks was subtracted from that evaluated on the target chunks. According to our model, if participants in the fast group learned more size three chunks than those in the accurate group, one would expect the fast group to have a higher measure of chunkiness than those in the accurate group.

Figure 4b shows the resulting chunkiness measure. The change Δ Wasserstein on size 3 chunks differed significantly between the two groups ($\chi^2(1) = 4.71$, $p = .02$), with the fast group showing a higher chunkiness compared to the accurate group ($\hat{\beta} = 12.33$, $t(88) = 2.15$, $p = .03$). Thus, participants in the fast condition showed a higher relative chunkiness in their reaction times to size three chunks in the sequence than participants in the accurate group.

Reaction Time Regression

Our model simulations showed that when the speed-accuracy trade-off parameter $w \to 0$ and accuracy becomes the only optimizing term, the model learns about the original transition matrix. However, as $w \to 1$ and speed is the only optimization term, the model learns a polarized transition probability where all the high and medium single element transitions become 1.

We aimed to test whether or not the w parameter of our model captured how the "fast" versus "accurate" instructions affected participants' chunking behavior. To this end, we fitted a linear mixed-effects regression to participants' log reaction times, with the group, the transition probabilities learned with $w = 0$ (true probabilities, Fig 2a left) and $w = 1$ (chunky probabilities, Fig 2a right) as independent variables, assuming a random intercept for each participant.

The best regression model contained the main effects of the transition probabilities learned with both values of w as well as two interaction effects with the given condition ($\chi^2(3) = 34.86$, $p < 0.001$). The first interaction was between the true probabilities and group (see Fig. 4d; $\hat{\beta} = -0.03$, $t(16740) = -2.74$, $p = 0.005$), as the true transition probabilities were more consistent with participants' responses in the accurate group than with those in the fast group. The second interaction was between the chunky probabilities and group ($\hat{\beta} = 0.05$, $t(16740) = 4.13$, $p < .001$), indicating that the simulated effect of a higher tendency to chunk was more predictive of the behavior of the fast group than that of the accurate group. Thus, the chunking bias induced by the w parameter of our model matched the bias observed in the reaction time pattern of the participants under speed demands.

Chunk Increase

As in Experiment 1, we again studied the increase in chunks ΔN from the baseline to the test blocks. Chunks were again identified by the mixture of Gaussians analysis. Shown in Figure 4e is the increase in the number of chunks for size 2 chunks (AB, BC, CD, DA) and size 3 chunks (chunk ABC, BCD, CDA, DAB) across the two groups.

Fitting a linear model on ΔN with group as the independent variable revealed a significant effect of group ($F(1) = 8.13$, $p = 0.005$). Compared to the accurate group, the fast group acquired more chunks from the baseline to the test block ($\hat{\beta} = 6.41$, $t(358) = 2.85$, $p = 0.005$). Thus, our model's prediction was confirmed.

Chunk Reuse

Finally, we compared the chunk reuse probability between the fast and the accurate group (see Figure 4f). Fitting a linear model to participants' chunk reuse probability showed that chunk reuse differed significantly between the two groups ($F(1) = 4.75$, $p = .03$). Participants in the fast group had a higher chunk reuse probability than participants in the accurate group ($\hat{\beta} = 0.07$, $t(114) = 2.18$, $p = .03$). This suggests that task demand modulated participants' chunk reuse. When participants tried to be faster, chunk reuse became more prominent.

Discussion

How people perceive and extract structure from a sequence of perceptual stimuli has been a longstanding question of psychological investigations. *Chunking* has been proposed as a mechanism to identify repeated patterns and segment sequences into those patterns. This way of segregating patterns into discrete chunks can improve storage, retrieval, and planning across multiple psychological domains.

In the current work, we have proposed that chunking benefits the timely and accurate execution of sequential actions. We used a rational model of chunking that adapts its representation to optimize a trade-off between speed and accuracy to simulate chunk learning in a serial reaction time task. Our simulations predicted that participants should chunk more if chunks are

indeed part of the generative model and should, on average, learn longer chunks when optimizing for speed than accuracy. We tested these predictions in two experiments. In Experiment 1, participants learned from sequences with different embedded chunks. In Experiment 2, participants were instructed to act as fast or accurately as possible. Multiple measures of chunking confirmed our model's predictions in both experiments. In sum, our results shed new light on the benefits of chunking and pave the way for future studies on step-wise representation learning in structured domains.

The model's prediction relating chunking to reaction time speed up relied partially on the Linear Ballistic Accumulator framework to translate within-chunk action prediction to an elevated starting point of the evidence accumulation, making the within-chunk action more likely to cross the decision threshold. Yet it remains challenging to explicitly fit a hierarchical LBA model over all participants, trials, and between-subject differences using our current data. This divergence is potentially due to a large number of observations. Therefore, one part of our analyses used model-predicted transition probabilities with accuracy and speed extremes to fit participants' reaction times. Nonetheless, future studies should look into the influence of chunking on the starting point of the LBA model in a fully Bayesian and hierarchically-structured model.

Currently, our model's predictions were primarily qualitative, and we did not compare across a more extensive set of alternative models. Even though we tested model-specific predictions such as the reuse of previously created chunks to parse the sequence and the speed-up and increased homogeneity of reaction times for within-chunk reaction times, future studies should further compare explicit predictions of different chunking models. We believe that our current work is a concrete first step towards building fine-grained models of human chunking in SRTs. We plan to compare our model to several alternatives in future tasks requiring participants to learn increasingly more hierarchically-structured chunks.

Finally, not all of our measures of increased chunking provided evidence for our model's predictions. In particular, in Experiment 1, the measure of chunkiness did not increase for the size 3 group even though we would have a priori expected such an increase. We believe that this increase did not appear because participants' speed-up of within-chunk reaction times was not uniform across both transitions of the size three chunks. Moreover, we did find a decrease of homogeneity for the size 2 group, which was as expected because learning about the size 2 chunk should makes the reaction time discrepancy between B and C larger. Importantly, we did find systematic differences across all other measures in both experiments and, therefore, believe that the current data support our model's predictions.

Related work

Our work can be related to several lines of previous research on chunking. Firstly, Servan-Schreiber and Anderson[27] studied how chunking facilitates memory by examining subjects' memorization for artificially produced grammatical sentences. They proposed that a hierarchy of chunks forms as subjects remember sentences. They instructed subjects to memorize sentences chunked by distinct hierarchy levels (e.g., word level vs. phrase level) and examined subjects' judgment of grammaticality afterwards. Their result suggested that the hierarchy of chunks influenced participants' grammaticality judgments. Apart from that, subjects overtly chunked the training sentences even when presented in an unstructured way. These findings are similar to our current model which also predicts chunking to appear in unstructured data but does so via a trade-off between accuracy and speed. The competitive chunking model provides a modeling framework consistent with our model but it does not explicate the processes that give rise to the construction of chunks and their hierarchy. The mechanism of recombining previously acquired chunks in our model can fill this blank.

Another related model is PARSER. Proposed by Perruchet and Vinter[28]. PARSER can produce artificial language stream segmentations of continuous input streams without any episodic cues such as pauses[29]. PARSER randomly samples the size of the next chunk of syllables and parses the sequence by disjunctive chunks. Each chunk learned by the model is associated with a weight, which increments with observational frequency and decrements via a mechanism of forgetting. Since both PARSER and our model evaluate chunks based on their occurrence statistics (PARSER approximates the chunk frequency online, whereas the rational chunking model evaluates the joint probability empirically), the simulation results produced by PARSER on syllable parsing can –in theory– be reproduced by our model. Distinct from PARSER and unique to our model is the mechanism of conjoining acquired chunks to construct new chunks and relating the general chunking mechanism to a rational form of utility maximization.

Wang et. al. trained a self-organized recurrent spiking neural network with spike-timing-dependent plasticity and homeostatic plasticity on sequences like the ones commonly used in SRTs, and showed that it could reproduce several sequence learning effects, in particular transfer effects[30]. It is, however, unclear whether the network learned explicit chunks that enabled this transfer because it is generally difficult to interpret the learned representations of such models. Compared to this approach, our model can serve as an interpretable computational level model because one can directly assess which chunks the model has learned. Apart from that, the simplicity of structure learning in our model allows neurally plausible implementations. We are, therefore, currently investigating how to implement our model in neuromorphic hardware, as well as using our model to reproduce transfer effects.

Another modeling approach to study how structure emerges from learning is to use variants of the Bayesian ideal observer

framework[31-33]. These models are also rational because their inference is evaluated on the observational instances. The difference between these models and our model mainly lies in the context window and their structural assumption. For example, with the hierarchical Dirichlet process model[?], the maximal size of the context window is pre-determined to evaluate the prediction of the next element given the previous context. In contrast, our model adapts its context length based on the previously acquired subsequences of chunks. Therefore, we think that these models are very similar to the accuracy part of our rational model and –in the limit– might even make the same predictions for bigrammatic chunks.

Finally, relying on the trade-off between speed and accuracy is only one way in which chunking can become beneficial, and other mechanisms such as minimizing memory or action complexity[34] have also been proposed. We believe that extending our current model to other domains with the by using these additional complexity measures will make scalable predictions spanning memory, reinforcement learning, and planning.

Conclusion

We investigated chunking behavior across two experiments and several measures. We found that chunking behavior depends on sequence statistics and task demands. When there are chunks in the training sequence, participants learn the underlying embedded chunks. Additionally, task demands modulate chunking behavior. Participants tend to chunk more when they are optimizing for speed rather than accuracy. Such chunking behavior occurs even in sequences lacking any deterministic transition probabilities. Our results suggest characteristics of chunking and how they interact with task demands. Our rational model of chunking captures and predicts these findings. The success of model predictions depends primarily on the gradual change in previously acquired representations to rationally adapt to sequence structure and task demands. We hope that our findings and model are a good step towards understanding human chunk learning across multiple domains.

Methods

Design
Experiment 1 was conducted using a between-groups designs in which participants were randomly assigned to one of three groups at the beginning of the experiment. These groups were the independent, size 2, and and size 3 conditions. The experiment was comprised of 10 blocks in total. The middle six blocks were the training blocks where participants practised the independent, size 2 or size 3 sequences. The first two and the last two blocks were the baseline and test blocks. In those blocks, all three groups of participants received the same instruction sequence generated from the "illusory" transition matrix. Training blocks differed between the three groups, as shown in Fig. 1d). The instruction sequence for the independent group was randomly and independently sampled from single-item elements A, B, C, and D with equal probability. This means that this instruction sequence contained no chunks. The instruction sequence for the size 2 group was generated by sampling AB, C, and D with an equal probability of 1/3. In other words, this instruction sequence contained the chunk AB. The instruction sequence for the size 3 was was generated by sampling ABC and D with an equal probability of 1/2. Thus, this instruction sequence contained the chunk ABC. All three groups received the same instruction to act "as fast and accurately as possible" throughout the experiment. On each trial, participants received feedback on the reaction time and correctness of the previous trial, followed by a 500ms response-to-stimulus interval. Participants were informed that their performance bonus on top of a base-pay was based on a mixture of their reaction times and accuracy.

Experiment 2 was also divided into 10 blocks with 100 trials each. Participants were randomly divided into two groups, creating a two-groups between-subjects design. One group was instructed to act as fast as possible (fast group) and one to act as accurate as possible (accurate group). Both groups were instructed to act "as fast and accurate as possible" in the first two and last two blocks (block 1-2 and 9-10, i.e. the baseline and test blocks, see Fig. 1a). In those blocks, accuracy and reaction times were displayed right at the end of each trial. In the middle 6 blocks, from block 3 to block 8 (i.e. the training blocks), participants in the fast group were instructed to act "as fast as possible even if it might lead to mistakes", and participants in the accurate group were instructed to act "as accurate as possible even it might slow you down". The fast group was told that their reward depended on how fast they pressed the instructed key and were given trial-by-trial feedback on their reaction times. The accurate group was told that their reward depended on their accuracy and were given trial-by-trial feedback on the correctness of their responses.

Recruitment of Participants
For Experiment 1, we recruited 142 participants from Amazon Mechanical Turk, out of which sixty-nine were female. Their median age was between 30 and 40, and the overall age ranged from 18 to above 50. This experiment took around 25 minutes to complete. After completing the task, participants received a base pay of $2 and a performance-dependent bonus of up to $6.

For Experiment 2, we recruited a total of 116 participants for our study, again from Amazon Mechanical Turk. Forty-eight participants were female; participants' median age was between 30 and 40, and the overall age ranged from 18 to above 50.

After completing the task, participants received a base pay of $2 and a performance-dependent bonus of up to $4.

Payment

For Experiment 1, a performance-dependent bonus was calculated as the weighted sum of participants' accuracy and reaction times. When the average accuracy was below 70%, the bonus was set to 0. The bonus for being fast was calculated as $bonusfast = bonusmax - (\overline{rt} - 600) * 0.025$, where $\overline{rt}$ indicates the average reaction time and $bonusmax$ indicates the maximal bonus. Participants were rewarded with a reaction time bonus when their average reaction time was below 600ms. Additionally, an accuracy bonus was calculated as the mean performance accuracy times the maximal bonus, $bonusacc = \overline{acc} \times bonusmax$. At the end of the experiment, the total bonus was calculated as a weighted average between the bonus for participants' accuracy and the bonus for their reaction times, $bonus = 0.5 \times bonusacc + 0.5 \times bonusfast$. If the final bonus was below 0, it was set to 0. If it was above the maximal bonus, it was set to the maximal bonus. On average, participants received $ 5.64 for their participation.

For Experiment 2, USD 2 were awarded as a base pay for every participant who completed the experiment. Additionally, participants received a performance-dependent bonus, ranging from 0 to the maximum of USD 4. This bonus was calculated separately for the fast and accurate groups. For the fast group, the bonus was 0 when their average accuracy was below 60%. If the average accuracy was above 60%, then the bonus was calculated as $bonusfast = (1000 - \overline{rt})/800 \times bonusmax$ This reward function penalized average reaction times that were slower than 1000ms. For the accurate group, the bonus was calculated as the percentage of their accuracy multiplied by the maximal bonus for the accuracy group, $bonusacc = \overline{acc} \times bonusmax$. Finally, the final bonus was again forcet to be between 0 and $bonusmax$ (USD 4). The mean reward earned by participants for this experiment was $4.16.

Filtering criteria

For Experiment 1, we excluded participants with an average reaction time longer than 1000ms or an average accuracy lower than 90%. Out of 142 participants who participated in Experiment 1, 20 participants were excluded and 122 remained after applying this exclusion criterion.

For Experiment 2, the same exclusion criteria were applied on the baseline and test blocks when both groups were asked to be as fast and accurate as possible. The exclusion criteria differed between the two groups on the training blocks. Participants in the fast group were excluded when the average reaction time was above 750ms ($n = 13$). Those in the accurate group were excluded when their accuracy was below 90% ($n = 10$). Additionally, we excluded participants who repeatedly failed attention checks before and after the experiment ($n = 3$). Out of 116 participants, 26 were excluded in total. All of the following analyses were performed on the data of the remaining 90 participants.

For the reaction-time based analysis including the chunky boost, chunkiness, and the mixture of Gaussians classification, we further excluded trials in which participants took more than 1000ms to respond. This amounted to 8.4% of all trials in Experiment 1, and 3.3% of all trials in Experiment 2.

Mixture of Gaussians Model

We used a mixture of Gaussians model to retrieve chunky transitions for each participant from their reaction times. To this end, the reaction time distribution for individual participants was used to classify individual trials as between or within-chunk reaction times. These reaction time distributions were fitted by a mixture of Gaussians model. The likelihood of belonging to the smallest mixture component was used to classify a reaction time as a within-chunk reaction time. This classification was then used for the identification of chunks for all experimental trials of every participant.

The classification of RTs by using multi-modal distributions was motivated by the idea that distinct processes might generate the within-chunk and between-chunk reaction times. During a SRT trial, if the participant has no expectation for the next upcoming instruction, she will first have to identify the instructed key on the screen before beginning to press a key. This will make her between-chunk reaction times larger. In contrast, if a participant has learned chunks, the she can expects the next instruction before it is even shown. If the upcoming instruction is within her expected chunk, the action to look for instructions displayed on the screen can be omitted, and she can directly engage in pressing their expected subsequent key. This will make her within-chunk reaction times smaller. Additionally, the mixture of Gaussians model also takes into account participants' post-error slow-downs. These correspond to the trials when a participant has made or almost made a mistake and corrects this tendency to press an expected key upon the observation of a conflicting instruction. The behavior of modifying the wrong key-press is slowing down the reaction time even more.

Because these three processes contribute to distinct components of participants' reaction times, a mixture of 3 Gaussian distributions was used to fit their reaction time distributions. We assumed that the within-chunk reaction time distributions had the lowest mean, the between-chunk reaction time distributions had a higher mean, and the post-error slow-down reaction time distribution had the highest mean. We fitted the mixture of Gaussian model to individual participants' reaction times, filtering

out RTs above 1000ms. A likelihood estimate belonging to each distribution amongst the mixture was assigned to the reaction times of each trial. A validation of this method has been included to the Supplementary Information.

Supporting information

Calculation of Chunky Boost
When the relevant chunk is AB, the reaction time is measured on B. The Cohen's d is measured on the speed-up between the baseline and the test block:

$$d_{B|A} = \frac{\bar{rt}_{baseline} - \bar{rt}_{test}}{\sigma} \tag{4}$$

σ is the pooled standard deviation based on the reaction time population of measured chunks.

The control reaction times came from size 2 subsequences that did not start with A, and the reaction times measured on the second item of the subsequence. Cohen's d was again evaluated as:

$$d_{\neg B|A} = \frac{\bar{rt}_{baseline} - \bar{rt}_{test}}{\sigma} \tag{5}$$

The chunky boost was calculated by subtracting the control from the relevant chunks:

$$\Delta d = d_{B|A} - d_{\neg B|A} \tag{6}$$

For Experiment 2, the relevant chunks are AB, BC, CD, and DA, the control subsequences were all size 2 subsequences excluding the relevant chunks.

Calculation of Chunkiness
When the relevant size 3 chunk was ABC, the Wasserstein distance was evaluated on rt_B and rt_C separately as $W(rt_B, rt_C)$. The chunkiness measure then was

$$\Delta W_{ABC} = W(rt_B, rt_C)_{baseline} - W(rt_B, rt_C)_{test} \tag{7}$$

The control chunks were subsequences that were not ABC, and their chunkiness measure became:

$$\Delta W_{control} = W(rt_2, rt_3)_{baseline} - W(rt_2, rt_3)_{test} \tag{8}$$

rt_2 and rt_3 denote the second and the third reaction time of the size 3 subsequences used as a control. The measure of chunkiness is the control subtracted from the relevant chunks.

$$Chunkiness = \Delta W_{ABC} - \Delta W_{control} \tag{9}$$

For Experiment 2, the relevant chunks were ABC, BCD, CDA, DAB, and the controls were all size 3 subsequences that were not any of the chunks. Chunkiness was evaluated in the same way.

Validation of mixture of Gaussian model
To perform a validation study on the mixture of Gaussians classification method, we simulated within- and between-chunk reaction time distributions by generating two exponentially modified Gaussian distribution (ex-Gaussian) with means apart. The ex-Gaussian distribution has the following probability density function.

$$f(x; \mu, \sigma, \lambda) = \frac{\lambda}{2} e^{\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)} erfc\left(\frac{\mu + \lambda\sigma^2 - x}{\sqrt{2}\sigma}\right) \tag{10}$$

erfc is the complementary error function defined as:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \tag{11}$$

μ is the mean of the independent Gaussian variable, σ is the standard deviation, and τ is the mean of the exponential component. The first three moments of the resulting ex-Gaussian distribution are $\mu + \tau$, $\sigma^2 + \tau^2$ and $2\tau^3$.

The exponentially modified Gaussian distribution normally provides a good fit for the reaction time distributions, as the simulated RT is straightly positive, and the exponential decay component takes account of the observed right-skewness of RT distributions[35,36].

To generate the distribution of within and between-chunk reaction times, we fixed the dispersion parameter (λ) to be 1. We varied the difference in the mean parameter μ between the two distributions, in addition to the spread σ of both distributions. After randomly interspersing samples coming from within and between-chunk distributions across 1000 trials (the same length as in the experiments), we used the mixture of three Gaussians to arrive at a classification accuracy. Plotted in Figure 5c is the classification accuracy of this method with increasing mean differences between the two distributions with different spread parameters σ. Note that for most of the parameters, the worst classification accuracy is still above 75%.

Mixture of Gaussian model comparison

When fitting the mixture of Gaussians model to individual participants' data, we separately varied the number of mixtures from 1 to 3 and used the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) to evaluate the three models. Shown in Figure 5b is the distribution of the best fitting mixtures over all participants. For most participants, the reaction time distribution was best described using 3 mixture components.
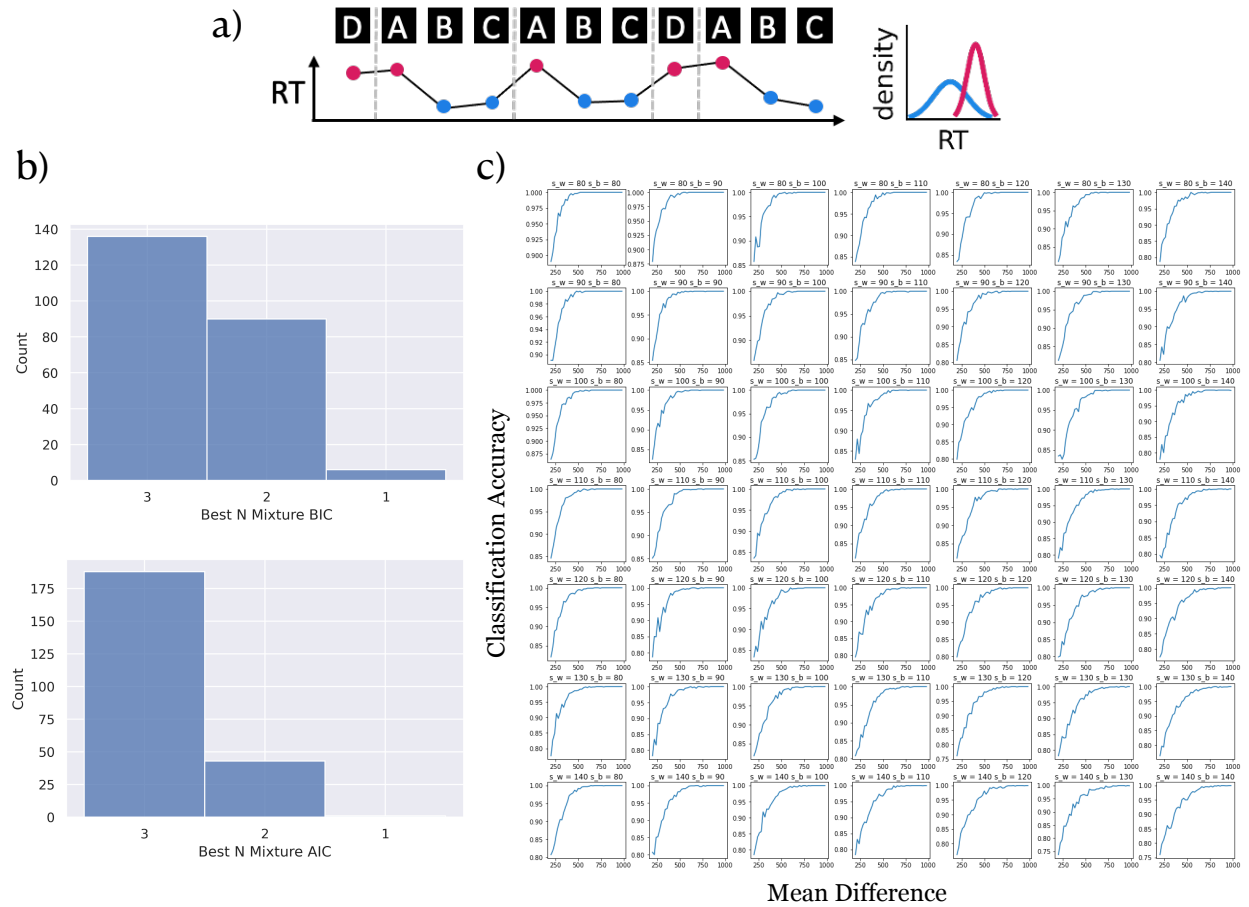

Figure 5. Mixture of Gaussians model for reaction time classification. a) Individual participant's reaction time of reacting upon key-press instructions across time. The distribution (right) is modelled as a mixture of two or three Gaussians. b) Histogram of the number of mixtures that led to best model selection criteria across all participants. Top: Best number of mixtures based on BIC (Bayesian Information Criteria). Bottom: Best number of mixtures based on AIC (Akaike Information Criteria). c) Classification accuracy (y-axis) of the mixture of Gaussians model on simulated reaction time data with increasing difference in mean (x-axis) and varying the standard deviation of the within-chunk (s_w) and between-chunk (s_b) distributions. s_w increases from the top to the bottom, and s_b increases from the left to the right.

Chunk Growth Rate Experiment 1

The mixture of Gaussian classification gives us the estimated chunking profile of every participant as reflected in their reaction time speed ups. We evaluated the growth rate of chunks on the three groups in Experiment 1 using this data. Fitting a linear mixed-effects regression onto participants' chunk size assuming random intercepts for each participant showed a significant effect of trial number ($\chi^2(1) = 51.22, p < 0.001$). Generally, the rate of increase was positive, suggesting that the chunk size acquired by participants grew as a function of time.

The chunk increase rate also differ significantly between the three conditions. The size 2 group ($\hat{\beta} = 1.43 \times 10^{-4}, t(84700) = 2.38, p = 0.02$) and the size 3 group had a significantly higher rate of chunk size increase compared to the independent group ($\hat{\beta} = 1.36 \times 10^{-4}, t(84700) = 2.93, p = 0.003$).

To summarize, the chunk size acquired by participants increased over time and differed across the three groups as predicted by our rational model of chunking.

Chunk Growth Rate Experiment 2

We evaluated the rate of chunk growth for Experiment 2 as well. Fitting a linear mixed-effects regression onto participants chunk size assuming random intercepts of participants also showed a significant effect of trial number ($\chi^2(1) = 778.82, p < 0.001$).

Additionally, there was an interaction effect between the groups and trial number. The fast group had a higher chunk learning rate than the accurate group ($\hat{\beta} = 2.68 \times 10^{-4}, t(62870) = 4.887, p < 0.001$). This suggests, as predicted, that the fast group had a higher tendency to build up chunks than the accurate group.

Acknowledgements

References

1. Miller, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* DOI: 10.1037/h0043158 (1956).

2. Laird, J. E., Rosenbloom, P. S. & Newell, A. Towards chunking as a general learning mechanism. In *AAAI*, 188–192 (1984).

3. Graybiel, A. M. The basal ganglia and chunking of action repertoires. *Neurobiol. learning memory* **70**, 119–136 (1998).

4. Servan-Schreiber, E. & Anderson, J. R. Learning artificial grammars with competitive chunking. *J. Exp. Psychol. Learn. Mem. Cogn.* **16**, 592 (1990).

5. Terrace, H. S. Chunking by a pigeon in a serial learning task. *Nature* **325**, DOI: 10.1038/325149a0 (1987).

6. Mathy, F. & Feldman, J. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* **122**, DOI: 10.1016/j.cognition.2011.11.003 (2012).

7. Lashley, K. S. *The problem of serial order in behavior*, vol. 21 (Bobbs-Merrill Oxford, United Kingdom, 1951).

8. Gobet, F. *et al.* Chunking mechanisms in human learning. *Trends Cogn. Sci.* **5**, DOI: 10.1016/S1364-6613(00)01662-4 (2001).

9. Graybiel, A. M. The basal ganglia and chunking of action repertoires. In *Neurobiology of Learning and Memory*, vol. 70, 1–2, DOI: 10.1006/nlme.1998.3843 (1998).

10. Egan, D. E. & Schwartz, B. J. Chunking in recall of symbolic drawings. *Mem. & Cogn.* **7**, DOI: 10.3758/BF03197595 (1979).

11. Ellis, N. C. Sequencing in SLA: Phonological memory, chunking, and points of order. *Stud. Second. Lang. Acquis.* **18**, DOI: 10.1017/S0272263100014698 (1996).

12. Koch, I. & Hoffmann, J. Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychol. Res.* **63**, DOI: 10.1007/PL00008165 (2000).

13. Brady, T. F., Konkle, T. & Alvarez, G. A. Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations. *J. Exp. Psychol. Gen.* **138**, DOI: 10.1037/a0016797 (2009).

14. Müssgens, D. M. & Ullén, F. Transfer in Motor Sequence Learning: Effects of Practice Schedule and Sequence Context. *Front. Hum. Neurosci.* **9**, DOI: 10.3389/fnhum.2015.00642 (2015).

15. Chase, W. G. & Simon, H. A. Perception in chess. *Cogn. Psychol.* **4(1)**, DOI: 10.1016/0010-0285(73)90004-2 (1973).

16. Gobet, F. & Simon, H. A. Expert chess memory: Revisiting the chunking hypothesis. *Memory* **6**, DOI: 10.1080/741942359 (1998).

17. Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. Compositional inductive biases in function learning. *Cogn. Psychol.* DOI: 10.1016/j.cogpsych.2017.11.002 (2017).

18. Schulz, E., Quiroga, F. & Gershman, S. J. Communicating compositional patterns. *Open Mind* **4**, 25–39 (2020).

19. Tomov, M. S., Yagati, S., Kumar, A., Yang, W. & Gershman, S. J. Discovery of hierarchical representations for efficient planning. *PLoS computational biology* DOI: 10.1371/journal.pcbi.1007594 (2020).

20. Nissen, M. J. & Bullemer, P. Attentional requirements of learning: Evidence from performance measures. *Cogn. psychology* **19**, 1–32 (1987).

21. Willingham, D. B., Nissen, M. J. & Bullemer, P. On the development of procedural knowledge. *J. experimental psychology: learning, memory, cognition* **15**, 1047 (1989).

22. Robertson, E. M. The serial reaction time task: implicit motor skill learning? *J. Neurosci.* **27**, 10073–10075 (2007).

23. Brown, S. & Heathcote, A. The simplest complete model of choice response time: linear ballistic accumulation. cognit. psychol. 57, 153-178. *Cogn. psychology* **57**, 153–78, DOI: 10.1016/j.cogpsych.2007.12.002 (2008).

24. Donkin, C., Brown, S. & Heathcote, A. Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *J. Math. Psychol.* **55**, 140–151 (2011).

25. Verwey, W. *et al.* Buffer loading and chunking in sequential keypressing. *J. Exp. Psychol.* **00**, 544–562, DOI: 10.1037//0096-1523.22.3.544 (1996).

26. Du, Y. & Clark, J. New insights into statistical learning and chunk learning in implicit sequence acquisition. *Psychon. Bull. Rev.* **24**, 1225–1233 (2017).

27. Servan-Schreiber, E. & Anderson, J. Learning artificial grammars with competitive chunking. *J. Exp. Psychol. Learn. Mem. Cogn.* **16**, 592–608, DOI: 10.1037/0278-7393.16.4.592 (1990).

28. Perruchet, P. & Vinter, A. Parser: A model for word segmentation. *J. Mem. Lang.* **39**, 246 – 263, DOI: https://doi.org/10.1006/jmla.1998.2576 (1998).

29. Saffran, J. R., Newport, E. L. & Aslin, R. N. Word segmentation: The role of distributional cues. *J. Mem. Lang.* **35**, 606–621, DOI: https://doi.org/10.1006/jmla.1996.0032 (1996).

30. Wang, Q., Rothkopf, C. A. & Triesch, J. A model of human motor sequence learning explains facilitation and interference effects based on spike-timing dependent plasticity. *PLoS computational biology* DOI: 10.1371/journal.pcbi.1005632 (2017).

31. Orbán, G., Fiser, J., Aslin, R. N. & Lengyel, M. Bayesian learning of visual chunks by human observers. *Proc. Natl. Acad. Sci. United States Am.* **105**, DOI: 10.1073/pnas.0708424105 (2008).

32. Goldwater, S., Griffiths, T. & Johnson, M. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* **112**, 21–54, DOI: 10.1016/j.cognition.2009.03.008 (2009).

33. O'Donnell, T. J., Goodman, N. D. & Tenenbaum, J. B. Fragment Grammars: Exploring Computation and Reuse in Language. Tech. Rep. (2009).

34. Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cognition* **204**, 104394 (2020).

35. Heathcote, A. Rtsys: A dos application for the analysis of reaction time data. *Behav. Res. Methods, Instruments, & Comput.* **28(3)**, 427–445, DOI: 10.3758/BF03200523 (1996).

36. Burbeck, S. L. & Luce, R. D. Evidence from auditory simple reaction times for both change and level detectors. *Percept. & Psychophys.* **32(2)**, 117–133, DOI: 10.3758/BF03204271 (1982).