When forgetting fosters learning: A neural network model for Statistical Learning

Ansgar D. Endress

Scott P. Johnson

Department of Psychology, City, University of London, UK

Department of Psychology, UCLA

Ansgar D. Endress
Department of Psychology
City, University of London
Northampton Square
London EC1V 0HB, UK
E-mail: ansgar.endress.1@city.ac.uk

Author Note

An *R* implementation of the model is available at `https://figshare.com/s/7a4ad045a3084f7b8920`. Please note that the URL will change in the final version of the manuscript. The final location will be `http://doi.org/10.25383/city.11359376`. This research was support by NIH grant R01-HD073535 to SPJ.

Abstract

Learning often requires splitting continuous signals into recurring units, such as the discrete words constituting fluent speech. A prominent candidate mechanism involves statistical learning of co-occurrence statistics like transitional probabilities (TPs), reflecting the idea that items from the same unit (e.g., syllables within a word) predict each other better than items from different units. TP computations are surprisingly flexible and sophisticated. Humans are sensitive to forward and backward TPs, compute TPs between adjacent items and longer-distance items, and even recognize TPs in novel units. We explain these hallmarks of statistical learning with a simple model with tunable excitatory connections and inhibitory interactions controlling the overall activation. With weak forgetting, activations are long-lasting, yielding associations among *all* items; with strong forgetting, no associations ensue as activations do not outlast stimuli; with intermediate forgetting, the network reproduces the hallmarks above. Forgetting thus is a key determinant of these sophisticated learning abilities.

Keywords: Statistical Learning; Implicit Learning; Transitional Probabilities; Neural Networks; Chunking

When forgetting fosters learning: A neural network model for Statistical Learning

1 Introduction

Observers often need to segment continuous signals into discrete recurring units, from the recognition of meaningful actions, where observers need to identify meaningful units in the continuous movement of other agents (Newtson, 1973; Zacks & Swallow, 2007) to language acquisition, where learners need to find out where words start and where they end in fluent speech (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). In the context of language acquisition, this challenge is called the segmentation problem (Aslin et al., 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) and is clearly one of the first challenges infants face, even before they can acquire the meaning of any word.

A prominent set of mechanisms for solving the segmentation problem relies on co-occurrence statistics of various sorts. These mechanisms track the predictability of items such as syllables. For example, predicting the next syllable after "the" is much harder than predicting the next syllable after "whis", because "the" can be followed by any noun while there are few possible continuations after "whis" (e.g., whiskey, whisker, ...). More formally, these predictive relationships have been quantified using Transitional Probabilities (TPs), i.e., the conditional probability of a syllable σ_2 following another syllable σ_1 $P(\sigma_2|\sigma_1)$.

After the initial discovery that infants and other animals are sensitive to TPs in general (Aslin et al., 1998; Chen & Ten Cate, 2015; Creel, Newport, & Aslin, 2004; Endress, 2010; Endress & Wood, 2011; Fiser & Aslin, 2002a; Hauser, Newport, & Aslin, 2001; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran & Griepentrog, 2001; Sohail & Johnson,

2016; Toro & Trobalón, 2005; Turk-Browne & Scholl, 2009), further research revealed the astonishing sophistication of these abilities.

For example, adults and infants can track backwards TPs (Endress & Wood, 2011; Perruchet & Desaulty, 2008; Pelucchi, Hay, & Saffran, 2009; Turk-Browne & Scholl, 2009) and discriminate high-TP items from low-TP items when the test-items are played in reverse order with respect to the familiarization (i.e., they readily recognize the item *CBA* after familiarization with *ABC*; Endress & Wood, 2011; Turk-Browne & Scholl, 2009). Learners can also track TPs between non-adjacent items (Endress, 2010; Endress & Wood, 2011; Peña, Bonatti, Nespor, & Mehler, 2002), though in some experiments, additional manipulations were required (Creel et al., 2004; Pacton & Perruchet, 2008). Both abilities are critical for language acquisition, because backwards TPs are in some languages more informative than forward TPs (e.g., Gervain & Guevara Erra, 2012) and because, across languages, non-adjacent dependencies abound (e.g., Newport & Aslin, 2004).

Learners prefer high-TP items to low-TP items even when the items are equated for frequency of occurrence (Aslin et al., 1998), and even when they had heard or seen only the low-TP items but not the high-TP items (Endress & Mehler, 2009; Endress & Langus, 2017; Perruchet & Poulin-Charronnat, 2012).

How can we make sense of these data? While a variety of computational models have been proposed to explain word segmentation (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Vinter, 1998), none of the extant models captures the sophistication of statistical learning abilities in their entirety.

For example, network models (such as Simple Recurrent Networks; Elman, 1990) are directional, and thus do not account for backward TPs, while their sensitivity to non-adjacent TPs will likely depend on the network parameters. "Chunking models" that store items in memory (Batchelder, 2002; Perruchet & Vinter, 1998) and information-theoretic models (or related Bayesian models) that minimize storage space in memory (Brent & Cartwright, 1996; Orbán et al., 2008) will not track (adjacent or non-adjacent) TPs in unattested items, and thus do not account for the entire range of data either.

Here, we suggest that an ability to succeed in the crucial test cases above follows naturally from a correlational learning mechanism such as Hebbian learning. Specifically, we assume that each item (syllable, visual shape, . . .) is represented by some population of neurons, and that participants are exposed to some sequence *ABCD*. . . , where each letter stands for an item. If the activation of such a population decays more slowly than the duration of an item, two adjacent items will be active simultaneously, and thus form an association. For example, if the representation of *A* is still active while *B* occurs, these representations will form an association. But if the representation of *A* is still active while *C* occurs, *A* and *C* will form an association as well even though they are not temporarily adjacent (see also Endress, 2010). Importantly, these associations are not directional: just as presenting *A* will activate *B*, presenting *B* will activate *A*.

Here, we provide a computational implementation of this model. The model is a fairly generic network, based on a widely used model of saliency maps in the parietal cortex to which we added a Hebbian learning component. We use this network architecture as it is fairly generic and widely used, but have no particular claims about attentional involvement in TP computations (but see e.g. Toro, Sinnett, & Soto-Faraco,

2005).

Specifically, the network consists of units that stand for populations of neurons encoding the items. Excitatory connections between units follow a Hebbian learning rule. To keep the total activation in the network at a reasonable level, we also added mutual interference among the units; the inhibitory interactions do not undergo learning.

Further specifics of the model can be found in Supplementary Information A.

2 Computational principles

We first illustrate the computational principles of the model by running a simulation with a stream consisting of 9 symbols *A*, *B*, ... *I* that are arranged into three three-item units *ABC*, *DEF* and *GHI*. Units were concatenated in random order so that each unit occurred 100 times.

Figure 1 shows the activation in response to the presentation of each item when the unit *ABC* is presented for the first time (a) and for the last time (b) as well as the weights between the underlying items after the last presentation.

Figure 1a shows that the *A* unit is still active when the *C* item is presented. As a result, we would expect a strong and reciprocal associative link between *A* and *B* and a weaker one between *A* and *C*, which is just what Figure 1c shows.

Comparing Figures 1a and b reveals that the activation of *A* is more reduced at its last occurrence. This is due to the inhibitory input from other units: On the first occurrence, no other units are active yet, and activation of *A* can only be reduced through inhibition when other units are active. In contrast, the activations of *B* and *C* do not seem reduced between Figures 1 a and c. This is because they receive excitatory input

from *A* (and *B* in the case of *C*) which compensates the inhibitory input from other units.

We will now use these computational principles to illustrate some of the critical results in the statistical learning literature.

3 Results

3.1 High- vs. low-TP items, tested forwards and backwards

We first explore the discrimination of high vs. low TP items after exposure to a sequence of 4 units of 3 items each (e.g., 4 words of 3 syllables). These units are randomly concatenated into a familiarization stream so that each unit occurs 100 times. We then present the network with test-items (see below) and record the total network activation while each item is presented. We hypothesize that the total activation provides us with a measure of the network's familiarity with the unit.[1]

This cycle of familiarization and test will be repeated 100 times, representing 100 participants.

While keeping the parameters for self-excitation and mutual inhibition constant (α and β in Supplementary Material A), we used forgetting rates (λ_a in Supplementary Material A) between 0 and 1. As forgetting in our model is exponential, a forgetting rate of zero means no forgetting, a forgetting rate of 1 implies the complete disappearance of activation on the next time step (unless a population of neurons receives excitatory input

[1] We also report simulations where we consider only those network activation in the items that are part of the current test-item rather than the global network activation. For example, when an unit *ABC* is presented, we assess the network's familiarity with the items by recording the activation in *A*, *B* and *C* – rather than the activation in *all* items. Intuitively, one would expect the results to be similar, as the active items will mainly be those that have been stimulated. These simulations are reported in Supplementary Information C.

from other populations), and a forgetting rate of .5 implies the decay of half of the activation.

3.1.1 Adjacent and non-adjacent forward TPs. We first evaluate the network's sensitivity to forward TPs among adjacent and non-adjacent items. These simulations are inspired by the paradigm by Saffran, Aslin, and Newport (1996) and Saffran, Newport, and Aslin (1996), among many others. After familiarization as described above, the network will be tested on units such as *ABC* and "part-units." Part-units are created either by taking the last two items from one unit and the first item from the next unit (e.g., *BC:D*, where the colon indicates the former unit boundary but is not present in the stimuli) or by taking the last item from one unit and the first two items from the next unit (e.g., *C:DE*). As a result, part-units have occurred during the familiarization sequence but straddled a unit boundary and thus have relatively weak TPs. We thus expect the network to be more familiar with units than with part-units.

The demonstration of a sensitivity to TPs among *non*-adjacent items is inspired by the paradigm by Endress and Bonatti (2007). Specifically, our high non-adjacent TP test-items take their first and the last item from the same unit, but the middle item from a different unit (e.g., *AGC*, where *A* and *C* come from the unit *ABC*, while *G* was the first item of the unit *GHI*). By analogy to Endress and Bonatti (2007), we call these items *rule-units*.

Our low non-adjacent TP test-items take their first and the last items from different units and take the middle item from yet another unit (e.g., *AGF*, where *A* is the first item from *ABC*, *F* is the last item from *DEF*, while *G* was the first item of the unit *GHI*). By analogy to Endress and Bonatti (2007), we call these items *class-units*. The critical difference between the rule-units and the class-units is that the TP between the first and the last item is 1.0 in rule-units and 0 in class-units.

We will also test a second rule-unit vs. class-unit contrast where the middle item is novel and did not appear in the familiarization stream (e.g., *AXC* vs. *AXF*, where *X* has never appeared in the familiarization stream).

For each comparison, we will create normalized difference scores to evaluate the model performance:

$$d = \frac{\text{Item}_1 - \text{Item}_2}{\text{Item}_1 + \text{Item}_2}$$

We then evaluate these difference scores against the chance level of zero using Wilcoxon tests. An alternative evaluation metric is to count the number of simulations (each representing a participant) preferring the target items, and to evaluate this count using a binomial test. With 100 simulations per parameter set, performance is significantly different from the chance level of 50% if at least 61 % of the simulations show a preference for the target items.

The results are shown in Figure 2a and 2b. For low forgetting rates (0 and 0.2), the network fails for all comparisons. This is unsurprising as low forgetting rates mean that all items remain active for many time steps, so that the network indiscriminately forms associations among virtually all items, and thus fails to track the statistical structure of the familiarization stream. Likewise, for the maximum forgetting rate, the network fails on all discriminations as well; this is again unsurprising, as no associations can be formed among items if forgetting is so strong that there is no overlap in activation between items.

Critically, for intermediate forgetting rates, the network performed well above chance for all comparisons. It performed somewhat better when contrasting units with

C:DE part-units, as has been observed in human participants by Fiser and Aslin (2002b). Importantly, however, all difference scores are clearly above chance, and between 83% and 100% of the simulations yielded positive difference scores (though only 63% yielded positive difference scores for forgetting rate .6 and non-adjacent TP comparisons). Further, adjacent TPs support higher forgetting rates than non-adjacent TPs, because activations need to last longer for non-adjacent TPs to be formed; while a sensitivity to TPs among adjacent items is maintained for a forgetting rate of 0.8, there is no such sensitivity to non-adjacent TPs.

3.1.2 Adjacent and non-adjacent backward TPS. There is considerable evidence that participants are not only sensitive to forward TPs, but also to backward TPs. They track TPs when the only informative TPs are backward rather than forward TPs (Perruchet & Desaulty, 2008; Pelucchi et al., 2009), and discriminate high-TP items from low-TP items when the test-items are played in reverse order (Endress & Wood, 2011; Turk-Browne & Scholl, 2009).

Here, we test the network's ability to track backward TPs by familiarizing the network with the same streams as in the previous section, but playing the test-items in reverse order (e.g., *CBA* instead of *ABC*).

As shown in Figure 3a and 3b, the network performance with reversed items essentially mirrors that with forward items, with similar performance for both forward and backward items, with the main difference that the performance asymmetry between *C:DE* and *BC:D* part-units was reversed.

3.2 The role of frequency of occurrence

The experiments presented so far confound TPs and frequency of occurrence: Units do not only have stronger TPs than part-units, but they also occur more frequently.

This problem was initially noted by Aslin et al. (1998). They addressed it by having infants "choose" between units and part-units that were matched in frequency (see Aslin et al., 1998 for more details on the design).

Endress and Mehler (2009) and Endress and Langus (2017) presented a more "extreme" control experiment. In their experiments, high-TP units were matched in terms of TPs to high-TP *phantom-units* that had the same TPs as units but never occurred in the familiarization stream and thus had a frequency of occurrence of zero (see Endress & Mehler, 2009; Endress & Langus, 2017 for more details on the design). Participants preferred (i.e., better recognized) high-TP units to low-TP part-units that had occurred in the familiarization stream, they preferred high-TP phantom-units to low-TP part-units despite the difference in frequency of occurrence, and they failed to discriminate between units and phantom-units (but see Perruchet & Poulin-Charronnat, 2012, for evidence that units and phantom-units might sometimes be discriminated).

Here, we expose the network to a six unit stream inspired by Endress and Mehler (2009) and Endress and Langus (2017). Following this, we test the network on units, phantom-units and part-units.

The results are shown in Figures 4a and 4b. As in the experiments reported above, the network failed on all comparisons for low forgetting rates as it indiscriminately learned associations among all items.

For medium and, in this experiment, high forgetting rates, the network preferred

units and, critically, also phantom-units over part-units roughly to the same extent; we also replicate the somewhat better performance when the part-unit is of *C:DA* type compared to part-units of *BC:D* type. As the participants in Endress and Mehler (2009) and Endress and Langus (2017), the network is thus more sensitive to differences in TPs than to differences in frequency of occurrence, and recognizes TPs even in items it has never encountered before.

In contrast, the network does not seem to discriminate between units and phantom-units, replicating Endress and Mehler's (2009) and Endress and Langus's (2017) results, and suggesting again that the network is more sensitive to TPs than to frequency of occurrence.

4 Discussion

Identifying recurrent units in a continuous signal is an important problem, especially for language acquisition. Observers might potentially solve this problem by tracking co-occurrence statistics among items, assessing the predictiveness of different items. Indeed, humans have sophisticated statistical learning abilities, allowing them to encode and recognize Transitional Probabilities (TPs) irrespective of whether items are played forward or backwards, whether the items are temporarily adjacent or non-adjacent, and whether the units in which the TPs occur are known or entirely novel.

We show that a simple neural network accounts for all of these phenomena based on correlational (i.e., Hebbian) learning. Interestingly, the critical ingredient for successful learning seems to be forgetting: If forgetting is too weak, indiscriminate associations are formed that are, therefore, uninformative; conversely, if forgetting is too strong, no associations are formed.

Our results also lead to a counterintuitive conclusion about the computational function of statistical learning. While our model presents a rather simple and straightforward mechanistic explanation for our sophisticated statistical learning abilities, these TP-based mechanisms are only partially compatible with the presumed function of statistical learning – namely to store recurrent units in memory. Ultimately, a mechanism that recognizes items played backwards or items it has not encountered at all can hardly be said to maintain faithful memory representations of the relevant items. Conversely, recognizing backwards or unheard items is inconsistent with models that actually store items in memory (Batchelder, 2002; Perruchet & Vinter, 1998).

Similar dissociations between statistical learning abilities and memory for specific episodes between amnesic and Parkinson's patients have led to the conclusion that humans have a (cortical) declarative memory system that is independent of a (neostriatal) system for forming associations (Knowlton, Mangels, & Squire, 1996; Poldrack et al., 2001). Statistical learning might be used for predictive processing rather than memory per se (Goujon, Didierjean, & Thorpe, 2015; Turk-Browne, Scholl, Johnson, & Chun, 2010), and our model is consistent with such a function.[2]

Together with our model, such results suggest that statistical learning, powerful as it is, might not be sufficient for placing recurring units in memory. After all, we clearly have declarative memories of such items, and know that we know the word *learning* rather than a backwards version such as *gninrael*. As a result, a critical question for future research is to find out how the power of predictive processes such as statistical

[2] While Parkinson's patients were initially thought to be impaired in associative learning in general (Knowlton et al., 1996), further research revealed that, for many tasks, such patients have intact associative learning abilities, and that their impairment might depend on the need to integrate probabilistic feedback across learning episodes (Smith & McDowall, 2006). Be that as it might, statistical learning does not seem to lead to declarative knowledge of specific even in studies that link it to the Medial Temporal Lobe (Turk-Browne et al., 2010).

learning is harnessed to form declarative memories of recurring units in sequences.

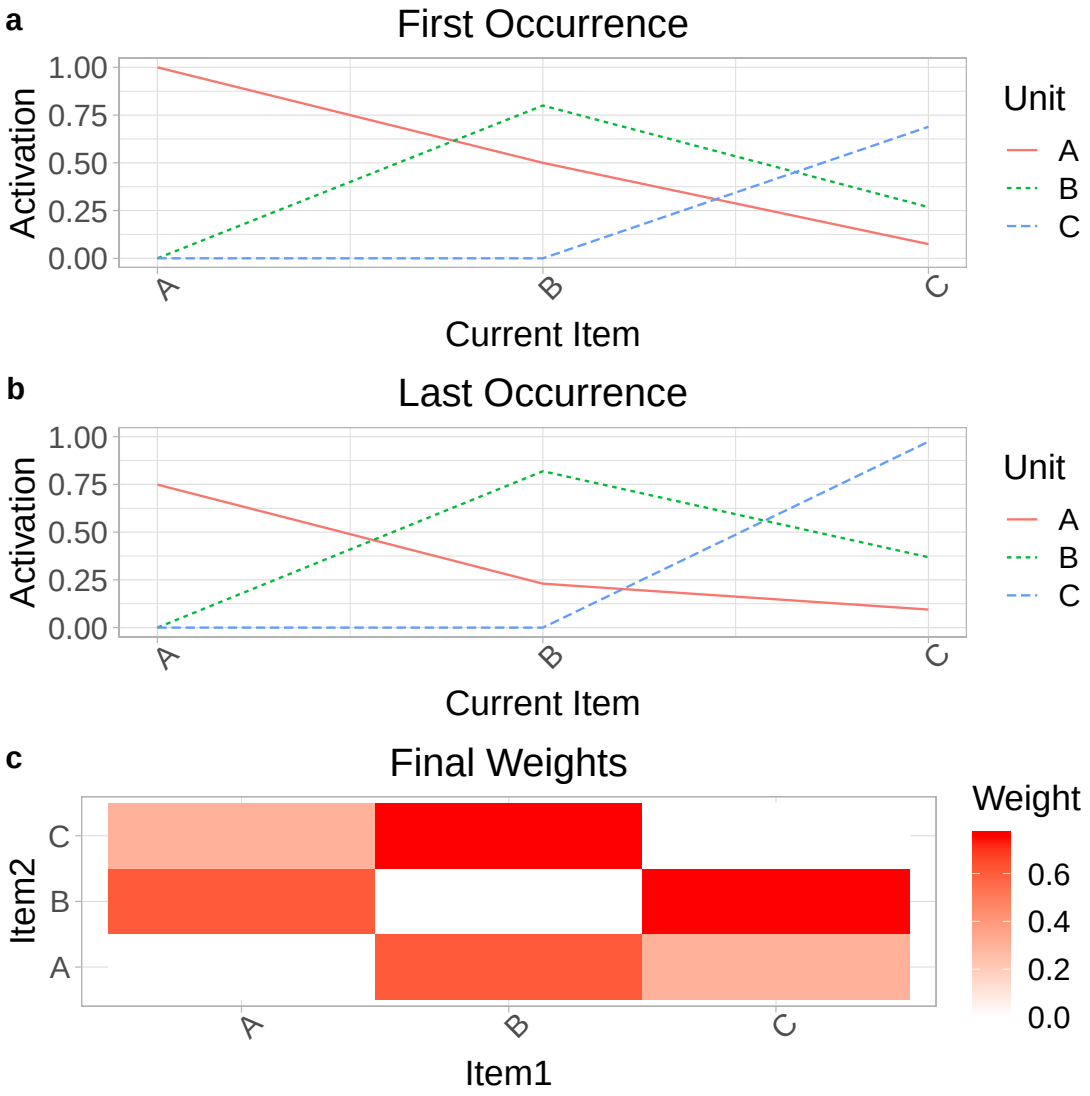

Figure 1. Illustration of the computational principles of the simulations. We plot the network activation when stimulated by a recurring unit *ABC*. (a) On the first occurrence of the unit, no associations have been formed yet. Hence, when A is presented, A (but no other items) becomes active, and then decays, though some activation persists even while C is presented. Likewise, B and C become active upon presentation, and then decay. The initial activation is weaker for B and C than for A due to the presence of inhibitory interactions; this is because, for A, no other potentially inhibiting representations are active yet, while other activated items (e.g., A) have inhibitory input for B and C. (b) On the last occurrence of a unit, associations between the items have been formed. When the network is externally stimulated with a unit such as *ABC*, the activation of B and C is greater than that of A when the corresponding items are stimulated. This is because B and C (but not A) receive excitatory input from the strongly associated, preceding items. (c) Weights at the end of the familiarization phase. The connection weights between adjacent items are stronger than those between non-adjacent items (i.e., between A and C).

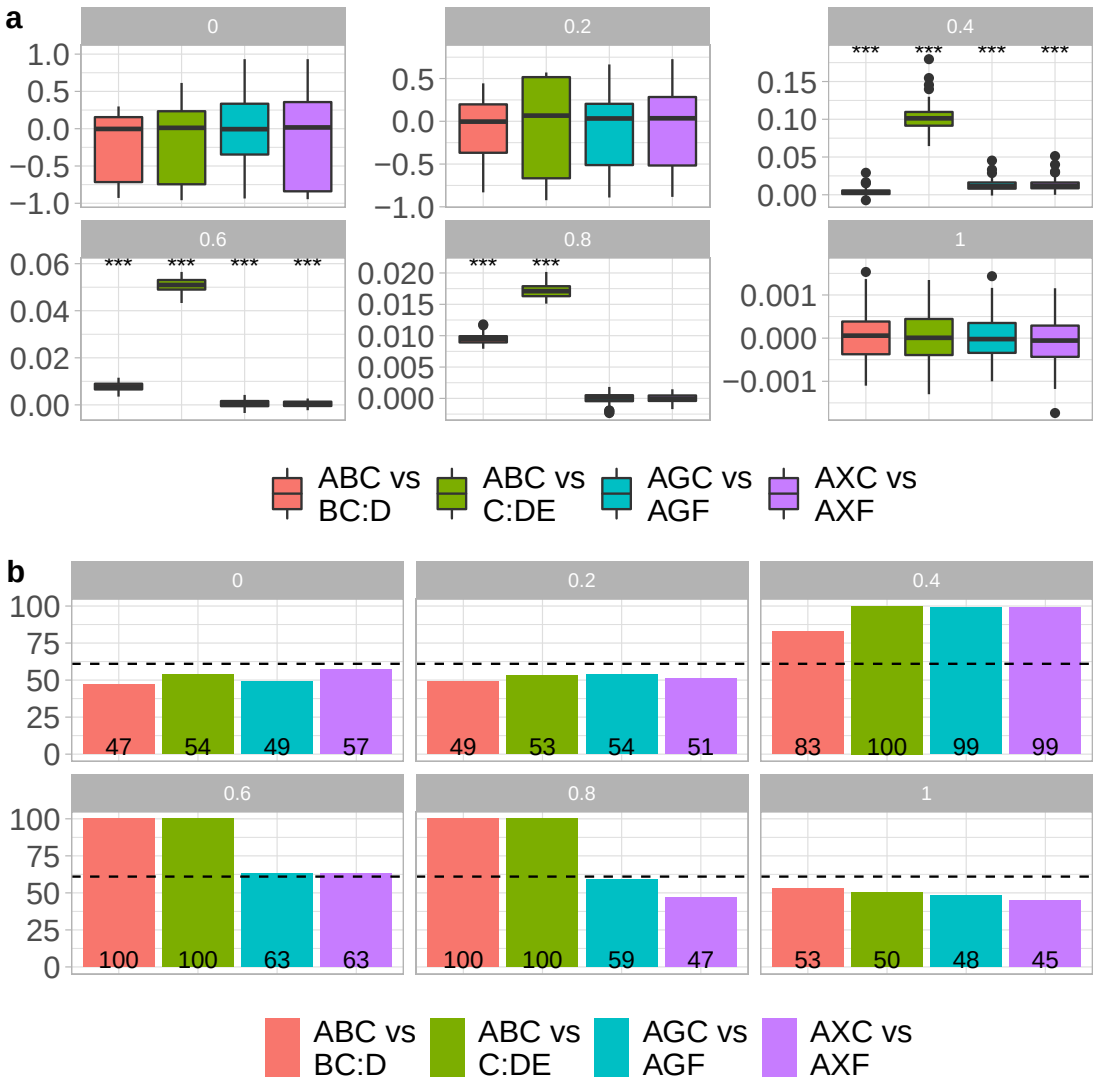

Figure 2. Results for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). (a) Difference scores. The scores are calculated based the global activation as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero. (b) Percentage of simulations with a preference for the target items. The simulations are assessed based on the global activation in the network. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

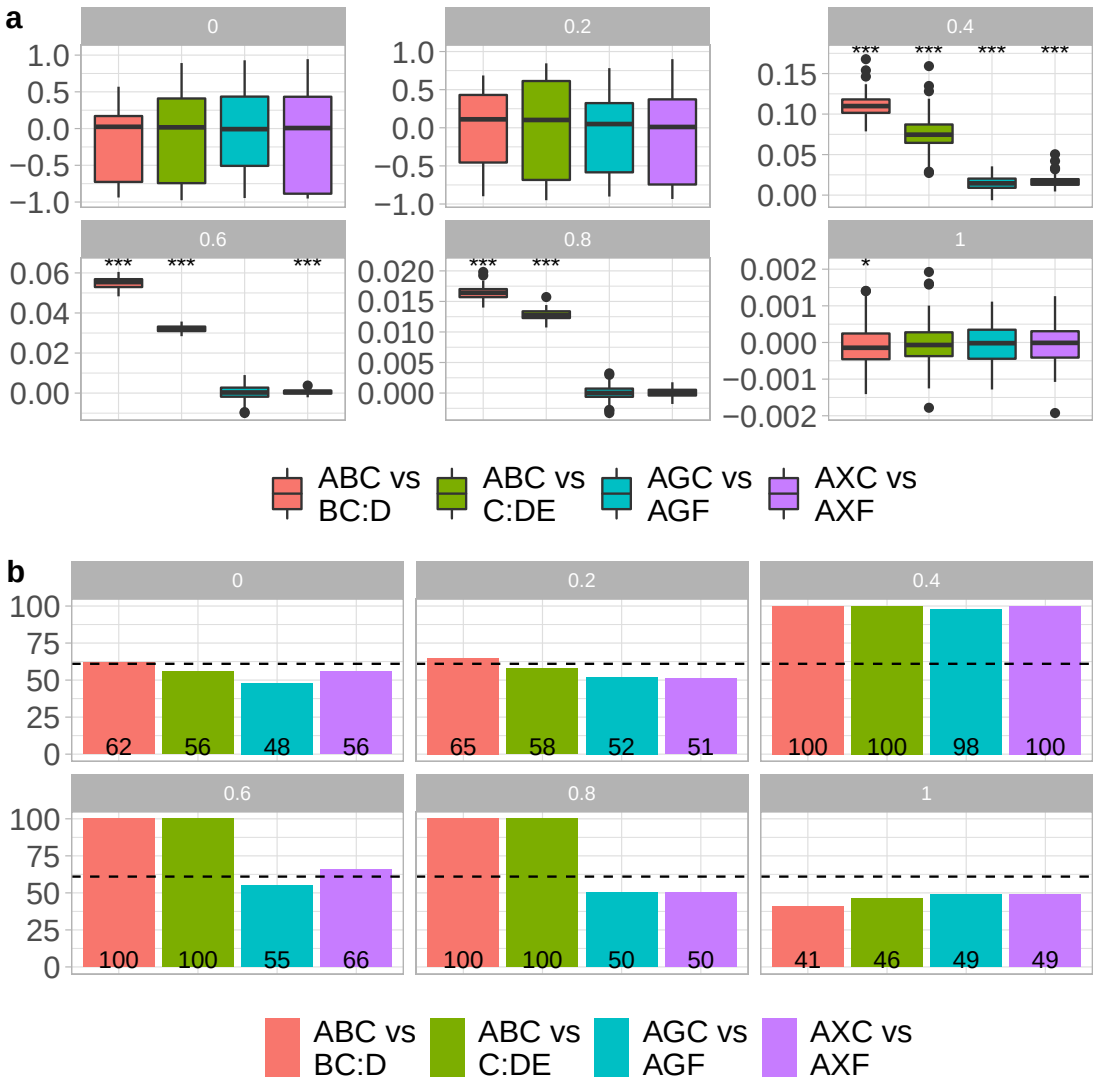

Figure 3. Results for items presented in **backward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). (a) Difference scores. The scores are calculated based the global activation as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero. (b) Percentage of simulations with a preference for the target items. The simulations are assessed based on the global activation in the network. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

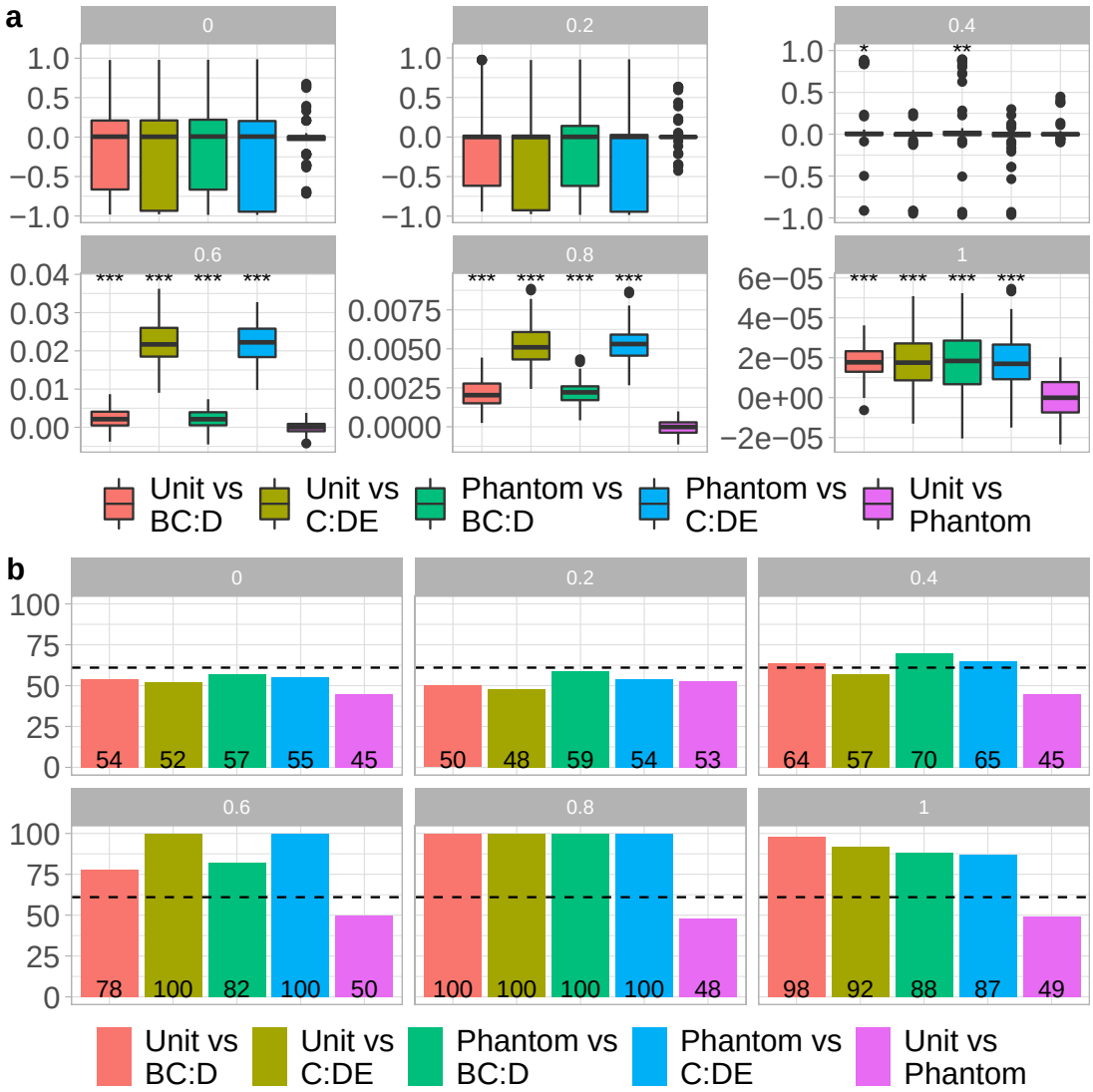

Figure 4. Results of the simulations comprising phantom-units, for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Phantom-Unit vs. Part-Unit: Phantom-Unit vs. *BC:D* and Phantom-Unit vs. *C:DE*; Unit vs. Phantom-Unit). (a) Difference scores. The scores are calculated based the global activation as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero. (b) Percentage of simulations with a preference for the target items. The simulations are assessed based on the global activation. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167-206.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93-125.

Chen, J., & Ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes*, *117*, 29–34. doi: 10.1016/j.beproc.2014.09.004

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(5), 1119-30. doi: 10.1037/0278-7393.30.5.1119

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–190.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. doi: 10.1016/j.cogpsych.2016.11.004

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When

fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351-367.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171.

Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(3), 458-67.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822-6. doi: 10.1073/pnas.232472899

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

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289-344.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: a study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287. doi: 10.1016/j.cognition.2012.06.010

Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in cognitive sciences*, *19*, 524–533. doi: 10.1016/j.tics.2015.07.009

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*, 1399–1402.

Letzkus, J. J., Wolff, S. B. E., Meyer, E. M. M., Tovote, P., Courtin, J., Herry, C., & Lüthi, A. (2011). A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature*, *480*, 331–335. doi: 10.1038/nature10674

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127-62.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology*, *28*(1), 28–38.

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

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(1), 80-96. doi: 10.1037/0278-7393.34.1.80

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-7. doi: 10.1016/j.cognition.2009.07.011

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-7. doi: 10.1126/science.1072901

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299–1305. doi: 10.3758/MC.36.7.1299

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. doi:

10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–63.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. doi: 10.1038/35107080

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-8.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74-85.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–21.

Smith, J. G., & McDowall, J. (2006). When artificial grammar acquisition in parkinson's disease is impaired: the case of learning via trial-by-trial feedback. *Brain research*, *1067*, 216–228. doi: 10.1016/j.brainres.2005.10.025

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 1-11. doi: 10.1080/15475441.2015.1073153

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25-34. doi: 10.1016/j.cognition.2005.01.006

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, *67*(5), 867-75.

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of neuroscience*, *30*, 11177–11187. doi: 10.1523/JNEUROSCI.0858-10.2010

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84.

Supplementary Material A

Model definition

The activation of the i^{th} unit $x_i(t)$ is governed by the differential equation.

$$\dot{x_i} = -\lambda_a x_i + \alpha \sum_{j \neq i} w_{ij} F(x_j) - \beta \sum_{j \neq i} F(x_j) + \text{noise}$$

where $F(x)$ is some activation function. (Here we use $F(x) = \frac{x}{1+x}$). The first term represents exponential forgetting with a time constant of λ_a, the second term activation from other units, and the third term inhibition among items to keep the overall activation in a reasonable range.

The weights w_{ij} are updated using a Hebbian learning rule

$$\dot{w}_{ij} = -\lambda_w w_{ij} + \rho F(x_i) F(x_j)$$

λ_w is the time constant of forgetting (which we set to zero in our simulations) while ρ is the learning rate.

A discrete version of the activation equation is given by

$$x_i(t+1) = x_i(t) - \lambda_a x_i(t) + \alpha \sum_{j \neq i} w_{ij} F(x_j) - \beta \sum_{j \neq i} F(x_j) + \text{noise}$$

While the time step is arbitrary in the absence of external input, we use the duration of individual units (e.g., syllables, visual symbols etc.) as the time unit in our

discretization because associative learning is generally invariant under temporal scaling of the experiment (Gallistel & Gibbon, 2000). Further, while only excitatory connections are tuned by learning in our model, the same effect could be obtained by tuning inhibition, for example through tunable disinhibitory interneurons (Letzkus et al., 2011). Here, we simply focus on the result that a fairly generic network architecture accounts for the hallmarks of statistical learning that, so far, have eluded explanation.

The discrete updating rule for the weights is

$$w_{ij}(t+1) = w_{ij}(t) - \lambda_w w_{ij}(t) + \rho F(x_i) F(x_j)$$

Simulation parameters are listed in Table A1. An *R* implementation is available at `https://figshare.com/s/7a4ad045a3084f7b8920`. (Please note that the URL will change in the final version of the manuscript. The final location will be `http://doi.org/10.25383/city.11359376`.)

Table A1
Parameters used in the simulations

Symbol	**Function**	**Value(s)**
α	Excitation coefficient	0.7
β	Inhibition coefficient	0.4
λ_a	Forgetting rate — Activation	0, 0.2, 0.4, 0.6, 0.8, 1
λ_w	Forgetting rate — Weights	0
$\sigma_{\text{noise, activation}}$	Standard deviation of activation noise	0.001
$\sigma_{\text{noise, weights}}$	Standard deviation of weight noise	0
ρ	0.05	

Supplementary Material B

Detailed results

Table B1 provides detailed results for the simulations in terms of descriptive statistics and statistical tests for the simulation testing the recognition of (forward and backward) units, part-units, rule-units and class-units.

Table B2 provides similar results for the simulations testing the recognition of units, part-units and phantom-units.

Table B1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the global activation as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores.*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
Forward					
0	*M*	-180×10^{-3}	-113×10^{-3}	-82.7×10^{-3}	-101×10^{-3}
0	*SE*	-18.1×10^{-3}	-11.4×10^{-3}	-8.31×10^{-3}	-10.2×10^{-3}
0	$p_{Wilcoxon}$	95.7×10^{-3}	222×10^{-3}	452×10^{-3}	607×10^{-3}
0	$P_{Simulations}$	470×10^{-3}	540×10^{-3}	490×10^{-3}	570×10^{-3}
200×10^{-3}	*M*	-109×10^{-3}	-72.8×10^{-3}	-92.6×10^{-3}	-87.1×10^{-3}
200×10^{-3}	*SE*	-11.0×10^{-3}	-7.32×10^{-3}	-9.31×10^{-3}	-8.75×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	120×10^{-3}	118×10^{-3}	152×10^{-3}	134×10^{-3}
200×10^{-3}	$P_{Simulations}$	490×10^{-3}	530×10^{-3}	540×10^{-3}	510×10^{-3}
400×10^{-3}	*M*	3.68×10^{-3}	102×10^{-3}	12.4×10^{-3}	13.2×10^{-3}
400×10^{-3}	*SE*	369×10^{-6}	10.2×10^{-3}	1.25×10^{-3}	1.33×10^{-3}

Table B1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the global activation as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
400×10^{-3}	$p_{Wilcoxon}$	2.92×10^{-12}	3.96×10^{-18}	4.08×10^{-18}	4.08×10^{-18}
400×10^{-3}	$P_{Simulations}$	830×10^{-3}	1.00	990×10^{-3}	990×10^{-3}
600×10^{-3}	*M*	7.65×10^{-3}	50.8×10^{-3}	565×10^{-6}	465×10^{-6}
600×10^{-3}	*SE*	769×10^{-6}	5.10×10^{-3}	56.8×10^{-6}	46.7×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	462×10^{-6}	320×10^{-6}
600×10^{-3}	$P_{Simulations}$	1.00	1.00	630×10^{-3}	630×10^{-3}
800×10^{-3}	*M*	9.48×10^{-3}	17.1×10^{-3}	-13.0×10^{-6}	-35.9×10^{-6}
800×10^{-3}	*SE*	953×10^{-6}	1.72×10^{-3}	-1.31×10^{-6}	-3.61×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	583×10^{-3}	681×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	590×10^{-3}	470×10^{-3}
1.00	*M*	32.9×10^{-6}	31.9×10^{-6}	23.7×10^{-6}	-64.9×10^{-6}
1.00	*SE*	3.30×10^{-6}	3.21×10^{-6}	2.38×10^{-6}	-6.52×10^{-6}
1.00	$p_{Wilcoxon}$	737×10^{-3}	646×10^{-3}	897×10^{-3}	231×10^{-3}
1.00	$P_{Simulations}$	530×10^{-3}	500×10^{-3}	480×10^{-3}	450×10^{-3}
Backward					
0	*M*	-125×10^{-3}	-82.7×10^{-3}	-79.9×10^{-3}	-74.8×10^{-3}
0	*SE*	-12.5×10^{-3}	-8.31×10^{-3}	-8.03×10^{-3}	-7.52×10^{-3}
0	$p_{Wilcoxon}$	947×10^{-3}	448×10^{-3}	286×10^{-3}	607×10^{-3}

Table B1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the global activation as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
0	$P_{Simulations}$	620×10^{-3}	560×10^{-3}	480×10^{-3}	560×10^{-3}
200×10^{-3}	*M*	9.35×10^{-3}	5.52×10^{-3}	-75.9×10^{-3}	-91.2×10^{-3}
200×10^{-3}	*SE*	940×10^{-6}	555×10^{-6}	-7.63×10^{-3}	-9.16×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	753×10^{-3}	730×10^{-3}	160×10^{-3}	92.4×10^{-3}
200×10^{-3}	$P_{Simulations}$	650×10^{-3}	580×10^{-3}	520×10^{-3}	510×10^{-3}
400×10^{-3}	*M*	111×10^{-3}	76.7×10^{-3}	14.9×10^{-3}	16.9×10^{-3}
400×10^{-3}	*SE*	11.2×10^{-3}	7.71×10^{-3}	1.50×10^{-3}	1.70×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	7.01×10^{-18}	3.96×10^{-18}
400×10^{-3}	$P_{Simulations}$	1.00	1.00	980×10^{-3}	1.00
600×10^{-3}	*M*	54.9×10^{-3}	32.2×10^{-3}	308×10^{-6}	536×10^{-6}
600×10^{-3}	*SE*	5.52×10^{-3}	3.23×10^{-3}	31.0×10^{-6}	53.9×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	239×10^{-3}	14.2×10^{-6}
600×10^{-3}	$P_{Simulations}$	1.00	1.00	550×10^{-3}	660×10^{-3}
800×10^{-3}	*M*	16.4×10^{-3}	12.8×10^{-3}	-22.4×10^{-6}	42.4×10^{-6}
800×10^{-3}	*SE*	1.65×10^{-3}	1.29×10^{-3}	-2.25×10^{-6}	4.26×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	985×10^{-3}	463×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	500×10^{-3}	500×10^{-3}
1.00	*M*	-118×10^{-6}	-50.9×10^{-6}	-47.2×10^{-6}	-22.9×10^{-6}

Table B1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the global activation as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
1.00	*SE*	-11.9×10^{-6}	-5.12×10^{-6}	-4.75×10^{-6}	-2.30×10^{-6}
1.00	$p_{Wilcoxon}$	39.6×10^{-3}	278×10^{-3}	358×10^{-3}	709×10^{-3}
1.00	$P_{Simulations}$	410×10^{-3}	460×10^{-3}	490×10^{-3}	490×10^{-3}

Table B2

Detailed results for the different forgetting rates and comparisons, using the global activation as a measure of the network's familiarity with the items. $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores.*

λ_a	Statistic	Unit vs BC:D	Unit vs C:DE	Phantom vs BC:D	Phantom vs C:DE	Unit vs Phantom
0	*M*	-57.8×10^{-3}	-121×10^{-3}	-49.7×10^{-3}	-91.3×10^{-3}	-38.7×10^{-3}
0	*SE*	-5.81×10^{-3}	-12.1×10^{-3}	-5.00×10^{-3}	-9.18×10^{-3}	-3.89×10^{-3}
0	$p_{Wilcoxon}$	876×10^{-3}	385×10^{-3}	865×10^{-3}	835×10^{-3}	133×10^{-3}
0	$P_{Simulations}$	540×10^{-3}	520×10^{-3}	570×10^{-3}	550×10^{-3}	450×10^{-3}
200×10^{-3}	*M*	-53.0×10^{-3}	-164×10^{-3}	-53.5×10^{-3}	-178×10^{-3}	27.6×10^{-3}
200×10^{-3}	*SE*	-5.33×10^{-3}	-16.5×10^{-3}	-5.38×10^{-3}	-17.8×10^{-3}	2.77×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	761×10^{-3}	120×10^{-3}	979×10^{-3}	111×10^{-3}	544×10^{-3}
200×10^{-3}	$P_{Simulations}$	500×10^{-3}	480×10^{-3}	590×10^{-3}	540×10^{-3}	530×10^{-3}
400×10^{-3}	*M*	76.4×10^{-3}	-27.0×10^{-3}	72.2×10^{-3}	-36.4×10^{-3}	14.3×10^{-3}
400×10^{-3}	*SE*	7.68×10^{-3}	-2.71×10^{-3}	7.25×10^{-3}	-3.66×10^{-3}	1.44×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	22.7×10^{-3}	819×10^{-3}	6.92×10^{-3}	471×10^{-3}	681×10^{-3}
400×10^{-3}	$P_{Simulations}$	640×10^{-3}	570×10^{-3}	700×10^{-3}	650×10^{-3}	450×10^{-3}
600×10^{-3}	*M*	2.06×10^{-3}	21.8×10^{-3}	2.12×10^{-3}	21.9×10^{-3}	-60.7×10^{-6}
600×10^{-3}	*SE*	207×10^{-6}	2.20×10^{-3}	214×10^{-6}	2.20×10^{-3}	-6.10×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	296×10^{-12}	3.96×10^{-18}	5.91×10^{-12}	3.96×10^{-18}	654×10^{-3}
600×10^{-3}	$P_{Simulations}$	780×10^{-3}	1.00	820×10^{-3}	1.00	500×10^{-3}
800×10^{-3}	*M*	2.12×10^{-3}	5.21×10^{-3}	2.17×10^{-3}	5.26×10^{-3}	-50.4×10^{-6}
800×10^{-3}	*SE*	213×10^{-6}	524×10^{-6}	218×10^{-6}	529×10^{-6}	-5.07×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	3.96×10^{-18}	3.96×10^{-18}	382×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	1.00	1.00	480×10^{-3}
1.00	*M*	17.8×10^{-6}	17.9×10^{-6}	17.5×10^{-6}	17.7×10^{-6}	233×10^{-9}
1.00	*SE*	1.79×10^{-6}	1.80×10^{-6}	1.76×10^{-6}	1.78×10^{-6}	23.4×10^{-9}
1.00	$p_{Wilcoxon}$	5.51×10^{-18}	172×10^{-18}	2.31×10^{-15}	846×10^{-18}	849×10^{-3}
1.00	$P_{Simulations}$	980×10^{-3}	920×10^{-3}	880×10^{-3}	870×10^{-3}	490×10^{-3}

Supplementary Material C

Experiments using the activation in the test-items

Here, we report on experiments where we evaluate the network performance using the activation of only those items that are part of the the test-items instead of the global activation. That is, when an unit *ABC* was presented, we assess the network's familiarity with the items by recording the activation in *A*, *B* and *C*; in contrast, in the simulation above, we recorded the activation in *all* items. Intuitively, one would expect the results to be similar, as the active items will mainly be those that have been stimulated.

C.1 High- vs. low-TP items, tested forwards and backwards

C.1.1 Adjacent and non-adjacent forward TPs. In this section, we seek to demonstrate that the network is sensitive to basic forward TPs among and non-adjacent items. Again, to demonstrate a sensitivity to TPs among adjacent items, the network will be tested on units and part-units. Likewise, the demonstration of a sensitivity to TPs among *non*-adjacent items is inspired by the paradigm by Endress and Bonatti (2007) and will be tested on rule-units vs. class-units, either with a middle item that appear during familiarization or with a novel middle item.

As shown in Figure C1 and C2, the results are very similar to those based on the global network activation reported above: The network fails for very low and very high forgetting parameters, and successeds on all comparisons with intermediate forgetting parameters. Numerically speaking, the results are similar to those used above as well.

C.1.2 Adjacent and non-adjacent backward TPS. Again, we test the network's ability to track backward TPs by familiarizing the network with the same streams as in the previous section, but playing the test-items in reverse order (e.g., *CBA*

instead of *ABC*).

As shown in Figures C3 and C4, the results are very similar to those based on the global network activation reported above: The network fails for very low and very high forgetting parameters, and successeds on all comparisons with intermediate forgetting parameters. Numerically speaking, the results are similar to those used above as well.

Table C1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the activation of the elements of the test-items as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores.*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
Forward					
0	*M*	-180×10^{-3}	-113×10^{-3}	-82.7×10^{-3}	-101×10^{-3}
0	*SE*	-18.1×10^{-3}	-11.4×10^{-3}	-8.31×10^{-3}	-10.2×10^{-3}
0	$p_{Wilcoxon}$	95.7×10^{-3}	222×10^{-3}	452×10^{-3}	607×10^{-3}
0	$P_{Simulations}$	470×10^{-3}	540×10^{-3}	490×10^{-3}	570×10^{-3}
200×10^{-3}	*M*	-109×10^{-3}	-72.8×10^{-3}	-92.6×10^{-3}	-87.1×10^{-3}
200×10^{-3}	*SE*	-11.0×10^{-3}	-7.32×10^{-3}	-9.31×10^{-3}	-8.75×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	120×10^{-3}	118×10^{-3}	152×10^{-3}	134×10^{-3}
200×10^{-3}	$P_{Simulations}$	490×10^{-3}	530×10^{-3}	540×10^{-3}	510×10^{-3}
400×10^{-3}	*M*	3.68×10^{-3}	102×10^{-3}	12.4×10^{-3}	13.2×10^{-3}
400×10^{-3}	*SE*	369×10^{-6}	10.2×10^{-3}	1.25×10^{-3}	1.33×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	2.92×10^{-12}	3.96×10^{-18}	4.08×10^{-18}	4.08×10^{-18}
400×10^{-3}	$P_{Simulations}$	830×10^{-3}	1.00	990×10^{-3}	990×10^{-3}

Table C1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the activation of the elements of the test-items as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
600×10^{-3}	*M*	7.65×10^{-3}	50.8×10^{-3}	565×10^{-6}	465×10^{-6}
600×10^{-3}	*SE*	769×10^{-6}	5.10×10^{-3}	56.8×10^{-6}	46.7×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	462×10^{-6}	320×10^{-6}
600×10^{-3}	$P_{Simulations}$	1.00	1.00	630×10^{-3}	630×10^{-3}
800×10^{-3}	*M*	9.48×10^{-3}	17.1×10^{-3}	-13.0×10^{-6}	-35.9×10^{-6}
800×10^{-3}	*SE*	953×10^{-6}	1.72×10^{-3}	-1.31×10^{-6}	-3.61×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	583×10^{-3}	681×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	590×10^{-3}	470×10^{-3}
1.00	*M*	32.9×10^{-6}	31.9×10^{-6}	23.7×10^{-6}	-64.9×10^{-6}
1.00	*SE*	3.30×10^{-6}	3.21×10^{-6}	2.38×10^{-6}	-6.52×10^{-6}
1.00	$p_{Wilcoxon}$	737×10^{-3}	646×10^{-3}	897×10^{-3}	231×10^{-3}
1.00	$P_{Simulations}$	530×10^{-3}	500×10^{-3}	480×10^{-3}	450×10^{-3}
Backward					
0	*M*	-125×10^{-3}	-82.7×10^{-3}	-79.9×10^{-3}	-74.8×10^{-3}
0	*SE*	-12.5×10^{-3}	-8.31×10^{-3}	-8.03×10^{-3}	-7.52×10^{-3}
0	$p_{Wilcoxon}$	947×10^{-3}	448×10^{-3}	286×10^{-3}	607×10^{-3}
0	$P_{Simulations}$	620×10^{-3}	560×10^{-3}	480×10^{-3}	560×10^{-3}
200×10^{-3}	*M*	9.35×10^{-3}	5.52×10^{-3}	-75.9×10^{-3}	-91.2×10^{-3}

Table C1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the activation of the elements of the test-items as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
200×10^{-3}	*SE*	940×10^{-6}	555×10^{-6}	-7.63×10^{-3}	-9.16×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	753×10^{-3}	730×10^{-3}	160×10^{-3}	92.4×10^{-3}
200×10^{-3}	$P_{Simulations}$	650×10^{-3}	580×10^{-3}	520×10^{-3}	510×10^{-3}
400×10^{-3}	*M*	111×10^{-3}	76.7×10^{-3}	14.9×10^{-3}	16.9×10^{-3}
400×10^{-3}	*SE*	11.2×10^{-3}	7.71×10^{-3}	1.50×10^{-3}	1.70×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	7.01×10^{-18}	3.96×10^{-18}
400×10^{-3}	$P_{Simulations}$	1.00	1.00	980×10^{-3}	1.00
600×10^{-3}	*M*	54.9×10^{-3}	32.2×10^{-3}	308×10^{-6}	536×10^{-6}
600×10^{-3}	*SE*	5.52×10^{-3}	3.23×10^{-3}	31.0×10^{-6}	53.9×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	239×10^{-3}	14.2×10^{-6}
600×10^{-3}	$P_{Simulations}$	1.00	1.00	550×10^{-3}	660×10^{-3}
800×10^{-3}	*M*	16.4×10^{-3}	12.8×10^{-3}	-22.4×10^{-6}	42.4×10^{-6}
800×10^{-3}	*SE*	1.65×10^{-3}	1.29×10^{-3}	-2.25×10^{-6}	4.26×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	985×10^{-3}	463×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	500×10^{-3}	500×10^{-3}
1.00	*M*	-118×10^{-6}	-50.9×10^{-6}	-47.2×10^{-6}	-22.9×10^{-6}
1.00	*SE*	-11.9×10^{-6}	-5.12×10^{-6}	-4.75×10^{-6}	-2.30×10^{-6}
1.00	$p_{Wilcoxon}$	39.6×10^{-3}	278×10^{-3}	358×10^{-3}	709×10^{-3}

Table C1
Detailed results for the different forgetting rates and comparisons (Unit vs. Part-Unit: ABC *vs.* BC:D *and* ABC *vs.* C:DE*; Rule-Unit vs. Class-Unit:* AGC *vs.* AGF *and* AXC *vs.* AXF*), for items presented in forward and backward order, and using the activation of the elements of the test-items as a measure of the network's familiarity with the items.* $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores. (continued)*

λ_a	Statistic	ABC vs BC:D	ABC vs C:DE	AGC vs AGF	AXC vs AXF
1.00	$P_{Simulations}$	410×10^{-3}	460×10^{-3}	490×10^{-3}	490×10^{-3}

C.2 The role of frequency of occurrence

As mentioned above, the experiments presented so far confound TPs and frequency of occurrence: Units do not only have stronger TPs than part-units, but they also occur more frequently. Among the control experiments for this issue (Aslin et al., 1998; Endress & Mehler, 2009; Endress & Langus, 2017), our computational experiments are inspired by Endress and Mehler (2009) and Endress and Langus (2017). We thus expose the network to a six unit stream inspired by Endress and Mehler (2009) and Endress and Langus (2017). Following this, we test the network on units, phantom-units and part-units.

As shown in Figure C5 and C6, the results are very similar to those based on the global network activation reported above: The network fails for very low and very high forgetting parameters, and prefers units and phantom-units over part-units roughly to the same extent for medium and high forgetting rates. As in Endress and Mehler (2009) and Endress and Langus (2017), it thus more sensitive to differences in TPs than to differences in frequency of occurrence. In contrast, the network does not seem to

discriminate between units and phantom-units, replicating Endress and Mehler's (2009) and Endress and Langus's (2017) results.

C.3 Detailed results

Table C1 provides detailed results for the simulations in terms of descriptive statistics and statistical tests for the simulation testing the recognition of (forward and backward) units, part-units, rule-units and class-units.

Table C2 provides similar results for the simulations testing the recognition of units, part-units and phantom-units.

Table C2

Detailed results for the different forgetting rates and comparisons, and using the activation of the elements of the test-items as a measure of the network's familiarity with the items. $p_{Wilcoxon}$ *represents the* p *value of a Wilcoxon test on the difference scores against the chance level of zero.* $P_{Simulations}$ *represents the proportion of simulations showing positive difference scores.*

λ_a	Statistic	Unit vs BC:D	Unit vs C:DE	Phantom vs BC:D	Phantom vs C:DE	Unit vs Phantom
0	***M***	-57.8×10^{-3}	-121×10^{-3}	-49.7×10^{-3}	-91.3×10^{-3}	-38.7×10^{-3}
0	***SE***	-5.81×10^{-3}	-12.1×10^{-3}	-5.00×10^{-3}	-9.18×10^{-3}	-3.89×10^{-3}
0	$p_{Wilcoxon}$	876×10^{-3}	385×10^{-3}	865×10^{-3}	835×10^{-3}	133×10^{-3}
0	$P_{Simulations}$	540×10^{-3}	520×10^{-3}	570×10^{-3}	550×10^{-3}	450×10^{-3}
200×10^{-3}	***M***	-53.0×10^{-3}	-164×10^{-3}	-53.5×10^{-3}	-178×10^{-3}	27.6×10^{-3}
200×10^{-3}	***SE***	-5.33×10^{-3}	-16.5×10^{-3}	-5.38×10^{-3}	-17.8×10^{-3}	2.77×10^{-3}
200×10^{-3}	$p_{Wilcoxon}$	761×10^{-3}	120×10^{-3}	979×10^{-3}	111×10^{-3}	544×10^{-3}
200×10^{-3}	$P_{Simulations}$	500×10^{-3}	480×10^{-3}	590×10^{-3}	540×10^{-3}	530×10^{-3}
400×10^{-3}	***M***	76.4×10^{-3}	-27.0×10^{-3}	72.2×10^{-3}	-36.4×10^{-3}	14.3×10^{-3}
400×10^{-3}	***SE***	7.68×10^{-3}	-2.71×10^{-3}	7.25×10^{-3}	-3.66×10^{-3}	1.44×10^{-3}
400×10^{-3}	$p_{Wilcoxon}$	22.7×10^{-3}	819×10^{-3}	6.92×10^{-3}	471×10^{-3}	681×10^{-3}
400×10^{-3}	$P_{Simulations}$	640×10^{-3}	570×10^{-3}	700×10^{-3}	650×10^{-3}	450×10^{-3}
600×10^{-3}	***M***	2.06×10^{-3}	21.8×10^{-3}	2.12×10^{-3}	21.9×10^{-3}	-60.7×10^{-6}
600×10^{-3}	***SE***	207×10^{-6}	2.20×10^{-3}	214×10^{-6}	2.20×10^{-3}	-6.10×10^{-6}
600×10^{-3}	$p_{Wilcoxon}$	296×10^{-12}	3.96×10^{-18}	5.91×10^{-12}	3.96×10^{-18}	654×10^{-3}
600×10^{-3}	$P_{Simulations}$	780×10^{-3}	1.00	820×10^{-3}	1.00	500×10^{-3}
800×10^{-3}	***M***	2.12×10^{-3}	5.21×10^{-3}	2.17×10^{-3}	5.26×10^{-3}	-50.4×10^{-6}
800×10^{-3}	***SE***	213×10^{-6}	524×10^{-6}	218×10^{-6}	529×10^{-6}	-5.07×10^{-6}
800×10^{-3}	$p_{Wilcoxon}$	3.96×10^{-18}	3.96×10^{-18}	3.96×10^{-18}	3.96×10^{-18}	382×10^{-3}
800×10^{-3}	$P_{Simulations}$	1.00	1.00	1.00	1.00	480×10^{-3}
1.00	***M***	17.8×10^{-6}	17.9×10^{-6}	17.5×10^{-6}	17.7×10^{-6}	233×10^{-9}
1.00	***SE***	1.79×10^{-6}	1.80×10^{-6}	1.76×10^{-6}	1.78×10^{-6}	23.4×10^{-9}
1.00	$p_{Wilcoxon}$	5.51×10^{-18}	172×10^{-18}	2.31×10^{-15}	846×10^{-18}	849×10^{-3}
1.00	$P_{Simulations}$	980×10^{-3}	920×10^{-3}	880×10^{-3}	870×10^{-3}	490×10^{-3}

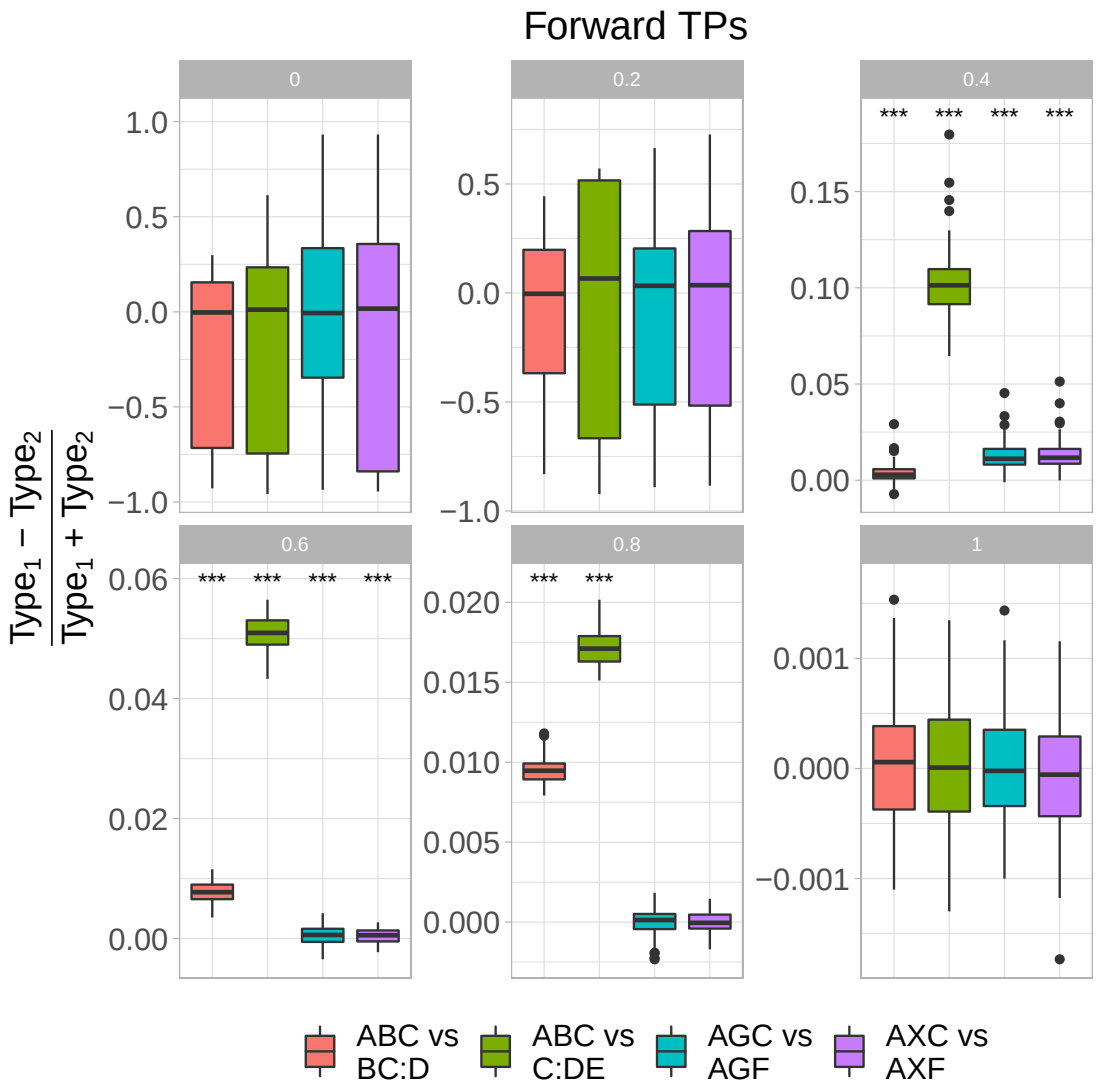

Figure C1. Difference scores for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). The scores are calculated based the activation in the test items as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero.

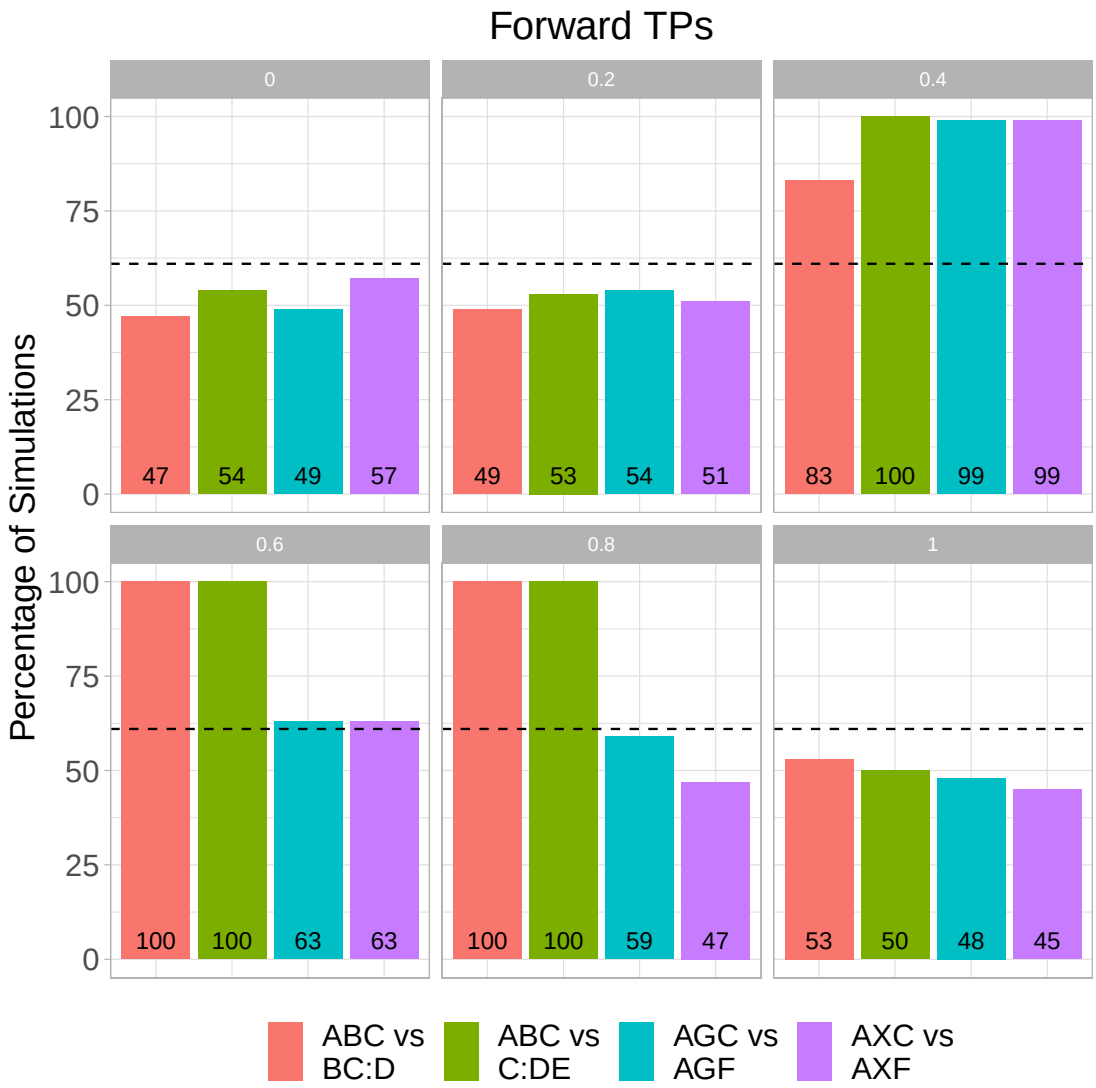

Figure C2. Percentage of simulations with a preference for the target items for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1) and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). The simulations are assessed based on the activation in the test items. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

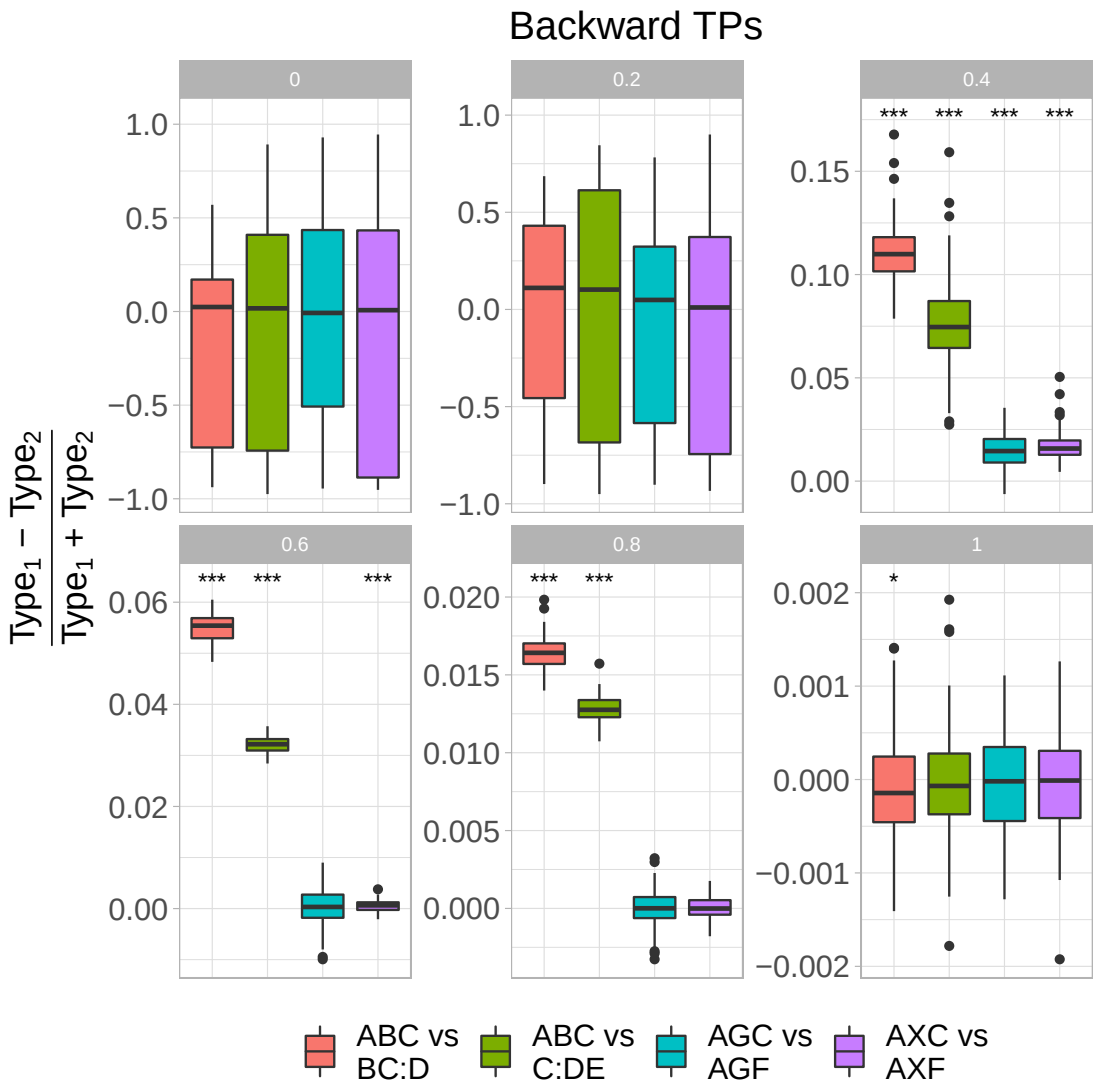

Figure C3. Difference scores for items presented in **backward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). The scores are calculated based the activation in the test items as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero.

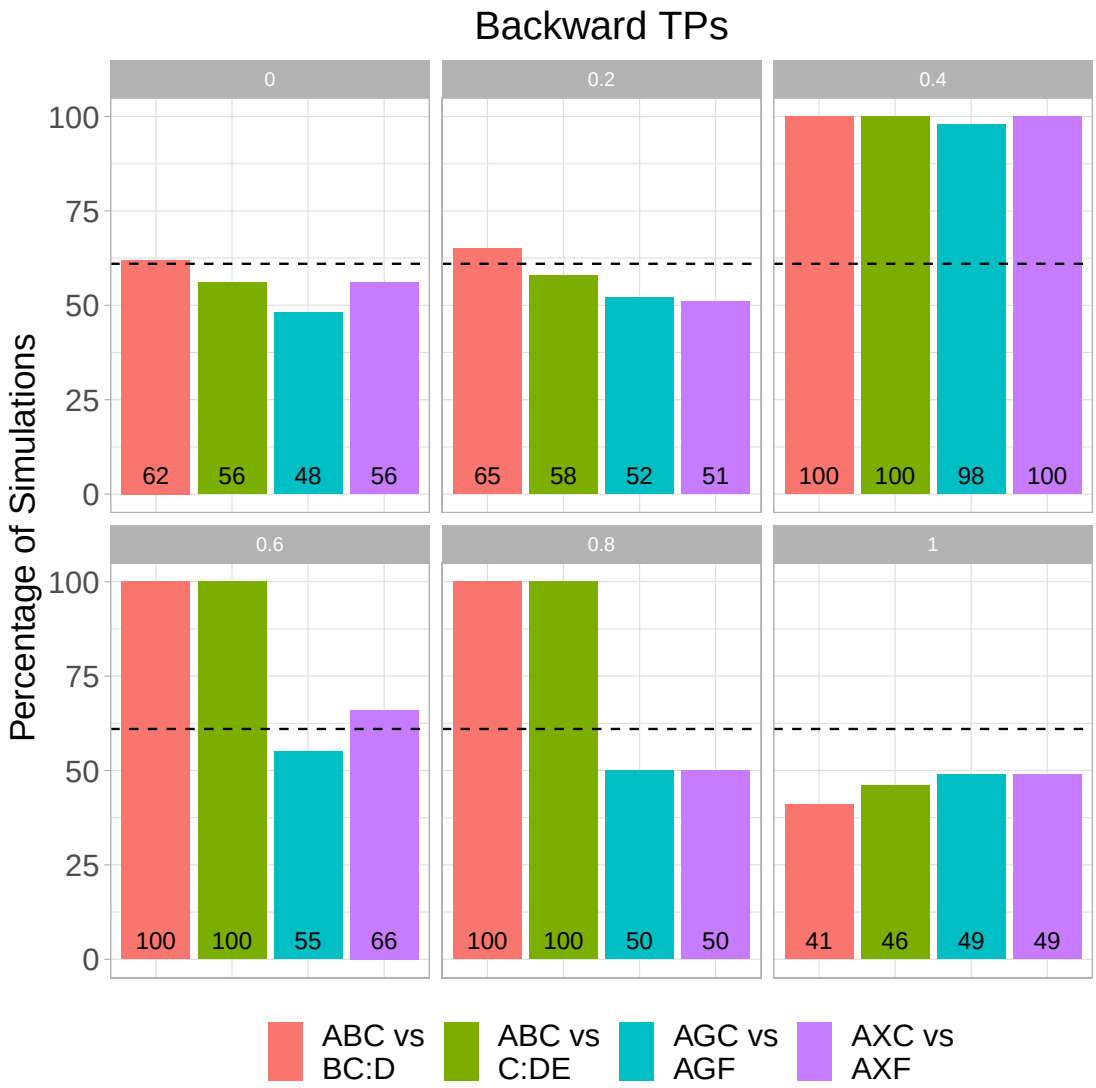

Figure C4. Percentage of simulations with a preference for the target items for items presented in **backward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1) and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Rule-Unit vs. Class-Unit: *AGC* vs. *AGF* and *AXC* vs. *AXF*). The simulations are assessed based on the activation in the test items. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.

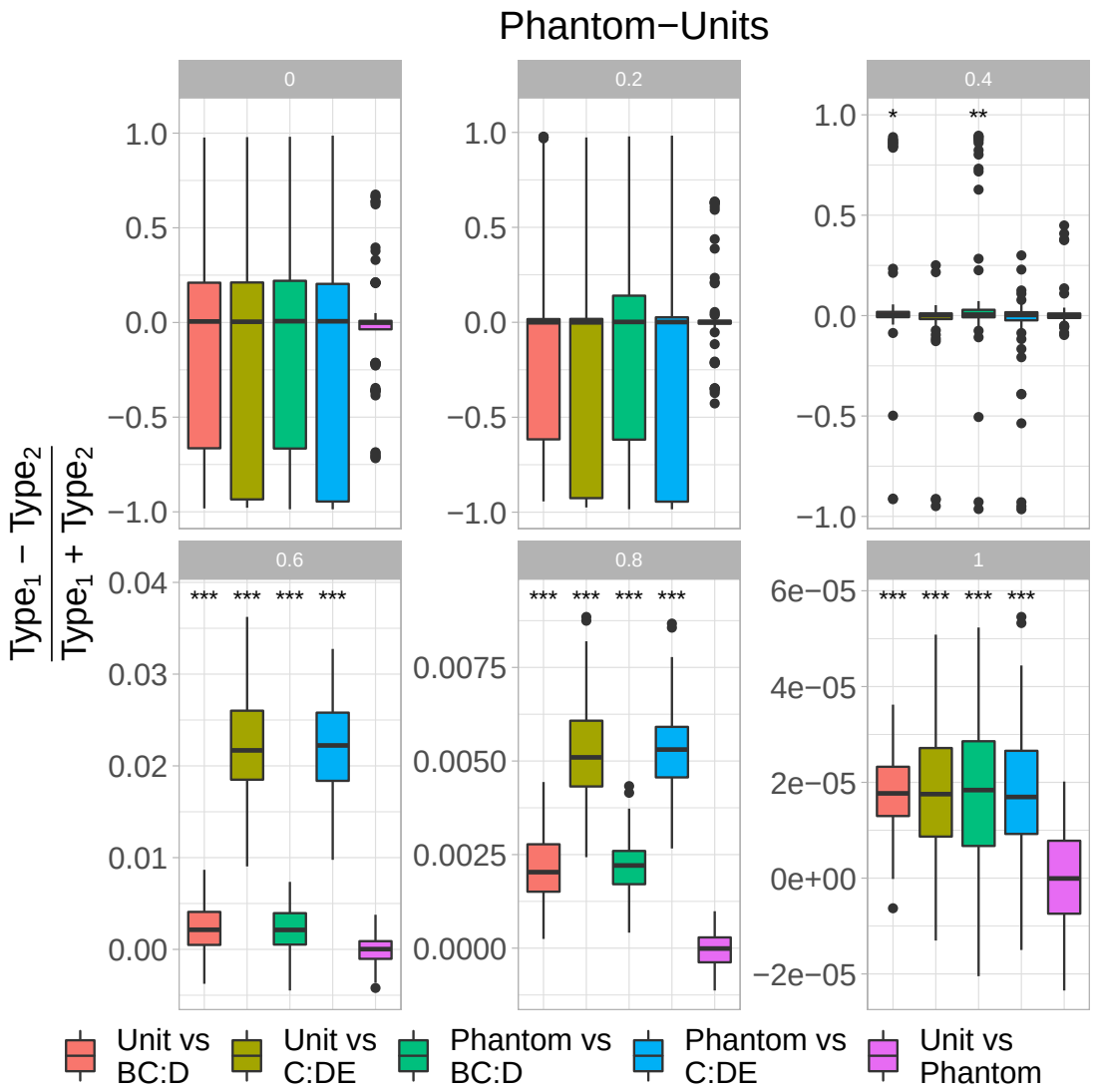

Figure C5. Difference scores for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1), and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Phantom-Unit vs. Part-Unit: Phantom-Unit vs. *BC:D* and Phantom-Unit vs. *C:DE*; Unit vs. Phantom-Unit). The scores are calculated based the activation in the test items as a measure of the network's familiarity with the items. Significance is assessed based on Wilcoxon tests against the chance level of zero.

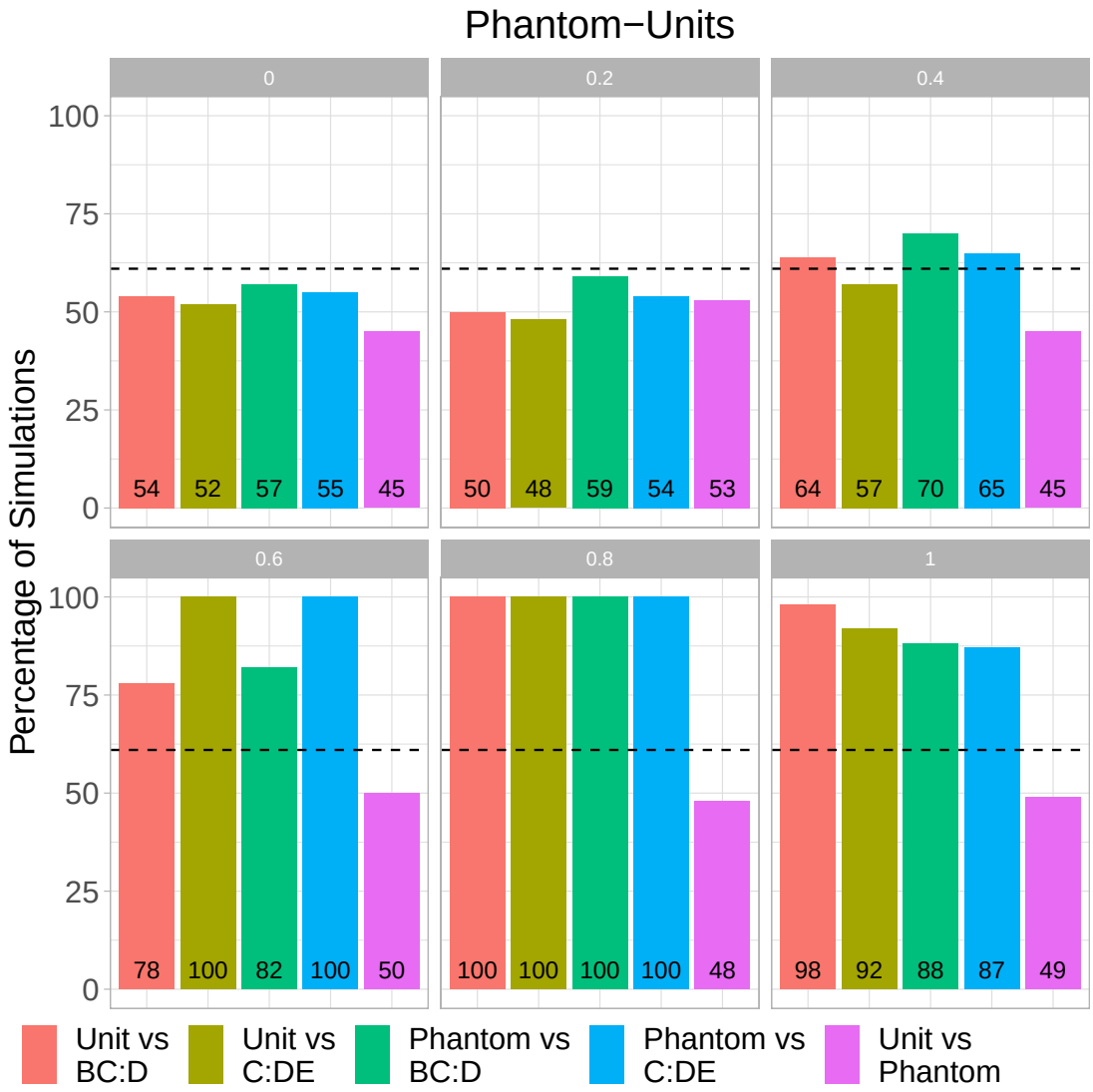

Figure C6. Percentage of simulations with a preference for the target items for items presented in **forward order**, different forgetting rates (0, 0.2, 0.4, 0.6, 0.8 and 1) and for the different comparisons (Unit vs. Part-Unit: *ABC* vs. *BC:D* and *ABC* vs. *C:DE*; Phantom-Unit vs. Part-Unit: Phantom-Unit vs. *BC:D* and Phantom-Unit vs. *C:DE*; Unit vs. Phantom-Unit). The simulations are assessed based on the activation in the test items. The dashed line shows the minimum percentage of simulations that is significant based on a binomial test.