Rule learning by Habituation can be Simulated in Neural Networks

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3A 1B1 Canada

Abstract

Contrary to a recent claim that neural network models are unable to account for data on infant habituation to artificial language sentences, the present simulations show successful coverage with cascade-correlation networks using analog encoding. The results demonstrate that a symbolic rule-based account is not required by the infant data.

One of the fundamental issues of cognitive science continues to revolve around which type of theoretical model better accounts for human cognition -- a symbolic rule-based account or a sub-symbolic neural network account. A recent study of infant habituation to expressions in an artificial language claims to have struck a damaging blow to the neural network approach (Marcus, Vijayan, Rao, & Vishton, 1999). The results of their study show that 7-month-old infants attend longer to sentences with unfamiliar structures than to sentences with familiar structures.

Because of certain features of their experimental design and their own unsuccessful neural network models, Marcus et al. conclude that neural networks cannot simulate these results and that infants possess a rule-learning capability unavailable to neural networks. A companion article suggests that rule learning is an innately provided capacity of the human mind, distinct from associative learning mechanisms like those in neural networks (Pinker, 1999).

My paper presents neural network simulations of the key features of the Marcus et al. (1999) experiment, thus showing that their infant data do not uniquely support a rule-based account.

Psychological Evidence and One Interpretation

Marcus et al. (1999) present experiments in which 7-month-old infants habituate to three-word sentences in an artificial language and are then tested on novel sentences that are either consistent or inconsistent with those to which the infant has habituated. In one experiment, illustrated in the first three columns of Table 1, infants habituated to sentences exhibiting an ABA pattern, for example, *ga ti ga* or *li na li*. There were 16 of these ABA sentences, created by combining four A words (*ga*, *li*, *ni*, and *ta*) with four B words (*ti*, *na*, *gi*, and *la*). Then the infants were presented with two novel sentences that were consistent with the ABA pattern (*wo fe wo*, and *de ko de*) and two novel sentences that were inconsistent with ABA because they followed an ABB pattern (*wo fe fe*, and *de ko ko*). A second, control condition habituated infants to sentences with an ABB pattern, for example, *ga ti ti* and *ga na na*. Again, 16 such sentences were created by combining the four A words with the four B words. The test sentences were the same in this second condition, but here the novel ABB sentences were consistent and the novel ABA sentences were inconsistent with the habituated ABB pattern.

Table 1: Conditions and error in simulation of Experiment 1

Procedure	Condition 1	Condition 2	Mean	SE
Habituate	ABA	ABB		
Consistent	ABA	ABB	0.649	0.107
Inconsistent	ABB	ABA	1.577	0.088

The dependent measure was looking time. During the test phase, if the infant looked at a flashing light to her left or right, a test sentence was played from a speaker near that light. A test sentence was played over and over until the infant either looked away or until 15 s elapsed. Infants attended more to inconsistent novel sentences than to consistent novel sentences, indicating that they were sensitive to grammatical differences between the sentences.

Marcus et al. designed another experiment, described in the first three columns of Table 2, that contrasted habituation to ABB sentences with AAB sentences. The idea was to rule out the possibility that infants might have used the presence or absence of duplicated words to distinguish grammatical types in their other experiments. For example, ABA sentences duplicate no words, but ABB sentences do (by duplicating B). In this Experiment 3, both grammatical sequences have duplicated words.

Table 2: Conditions and error in simulation of Experiment 3

Procedure	Condition 1	Condition 2	Mean	SE
Habituate	ABB	AAB		
Consistent	ABB	AAB	0.570	0.100
Inconsistent	AAB	ABB	1.491	0.072

Infants performed in a similar fashion in both experiments, i.e., they attended more to inconsistent than to consistent novel sentences. All infants except one showed the predicted preference for inconsistent over consistent test sentences. The issue is the proper theoretical account of this grammatical knowledge -- is it based on rules or on connections?

Marcus et al. argue that these simple grammars could not be learned by a computational system that is sensitive only to transitional probabilities or event frequencies.

Transitional probabilities would not work because the transitional probabilities for novel words would be 0. Counting the numbers of duplicated words might work for Experiment 1, but not for Experiment 3, where both grammars had duplicate words. Nonetheless, Marcus et al. briefly mention their unsuccessful attempts to simulate these habituation data with simple recurrent networks such as those used by Elman (1990).

No details of the course of habituation of attention or the extent of recovery were reported by Marcus et al. Nor was there an implementation of a rule-based model to account for the habituation data or a theoretical analysis of how rule learning might be used in the computation of habituation. In any case, the challenge raised by Marcus et al. is interesting and worthwhile. It is interesting because habituation is important and still poorly understood, and worthwhile because of the implications for the fundamental debate on rules vs. connections.

Habituation as Encoding and Decoding in Neural Networks

One computational account of habituation has been in terms of the encoding and decoding processes involved in so-called encoder networks (Mareschal & French, 1997). Encoder networks have output units identical to their input units. Their task is to reproduce their inputs on their output units. With layer-to-layer connectivity, an encoder network must encode input signals onto a layer of hidden units and then decode the hidden unit representations onto the output units. If the number of hidden units is less than the number of input or output units, then the encoder network learns to abstract a compact representation of the problem on its hidden units. Such compact abstractions generalize to novel inputs and enable prototype phenomena and pattern completion skills (Hertz, Krogh, & Palmer, 1991).

How might encoder networks be related to habituation? The habituation technique is arguably the most important methodological advance in developmental psychology in this century. The reason for this is that habituation enables the systematic study of perceptual and conceptual abilities in non-verbal, response-impoverished infants (Cohen, 1979). Unlike the study of mere preferences, habituation can be used even when no preferences exist. Even if the infant exhibits no natural preferences between stimulus categories, such preferences can be experimentally introduced by habituating the infant to one category and measuring dishabituation to another contrasting category. The responses required to show habituation and dishabituation are available at birth (Slater, 1995). All that is required is visual attention, head turning, or something as passive as heart rate. These advantages have enabled dozens of discoveries of perceptual and cognitive abilities in young infants over the past 30 years using habituation methodology. Infants have been demonstrated to perceive color, form, complex patterns, faces, and intricate relations, to learn categories and prototypes, to perceive perceptual constancies, to know about object permanence and causality, to identify objects, and to form both short-term and long-term memories for objects and events (e.g., Cohen, 1979; Haith, 1990; Oakes & Cohen, 1990; Quinn & Eimas, 1996). The memories identified in habituation studies are essentially recognition memories -- the recognition of a stimulus as being a member of a previously habituated category.

What is going on during the processes of habituation and dishabituation? The standard view is that infants gradually construct representational categories for stimuli that they encounter (Cohen, 1973; Sokolov, 1963). This category building is enabled by visual attention, as well as by other sensory modalities. Once a representational category is constructed via attention and processing, the infant no longer needs to attend so much to stimuli of that category. When the infant encounters a new stimulus, he compares it to stored representations of existing stimulus categories. If the new stimulus matches a stored category, then it will likewise elicit little or no attention. But if the new stimulus is not recognized as a member of an existing category, then it receives additional attention and processing. This is a system that seems adaptive in encouraging the infant to expend cognitive resources on novel information and thus continue to learn about the world.

There are many interesting aspects to the habituation literature. Among them is a tendency for attention to habituate gradually in a negatively accelerated fashion -- fast at first and then slowing down to an asymptote of no attention. The gradual decrease is perhaps a natural consequence of the fact that building representations in relatively naive infants takes time and effort. The negative acceleration is a natural consequence of the fact that attention to a stimulus may start at a high level and is bounded at none.

The basic idea enabling a link between habituation and encoder networks is that encoder networks model how one might learn about stimuli from attending to them. Relations among stimulus features are abstracted in the hidden unit representations as connection weights are adjusted. New stimuli that produce similar representations and little or no error are in a sense recognized as familiar. Those stimuli that produce different representations and large error are essentially unrecognized and considered as novel. The key assumption in this modeling of habituation with encoder networks is that network error corresponds to the need to direct current attentional resources (Mareschal & French, 1997). The theoretical contribution of this analysis is a computationally precise implementation of the representation and processing involved in habituation as presently understood.

The Current Model

Cascade-correlation

The proposed model of habituation uses an encoder version of the cascade-correlation learning algorithm. Cascade-correlation is an algorithm for learning in feed-forward neural networks (Fahlman & Lebiere, 1990). Unlike

standard back-propagation networks, whose topologies are designed by hand and remain static as connection weights are adjusted, cascade-correlation networks grow as well as learn. They grow by recruiting new hidden units into the network as required to reduce error at the output units. New hidden units are recruited one at a time and installed each on a separate layer with input connections from the input units and from any existing hidden units. The candidate hidden unit that gets recruited is the one whose activations correlate most highly with the network's current error as the input weights to the candidates are adjusted.

Cascade-correlation also differs from standard back-propagation by using curvature as well as slope information from the error surface in making weight adjustments. This additional information about the error surface, which is approximated in a computationally efficient way, enables more decisive and effective weight adjustments.

Cascade-correlation was designed to solve two of the major problems with back-propagation -- slow learning and inability to learn some difficult problems. On average, it learns about 10-50 times faster than standard back-propagation, and it learns problems that are too difficult for standard back-propagation networks (Fahlman & Lebiere, 1990). Some of the neurological justification for generative networks such as cascade-correlation is reviewed by Quartz and Sejnowski (1997).

Cascade-correlation has proved useful in simulating many aspects of cognitive development, including the balance scale (Shultz, Mareschal, & Schmidt, 1994), conservation (Shultz, 1998), seriation (Mareschal & Shultz, in press), pronoun semantics (Takane, Oshima-Takane, & Shultz, 1995), number comparison (Hashmi & Shultz, 1998), discrimination shift learning (Sirois & Shultz, 1998), and integration of velocity, time, and distance cues (Buckingham & Shultz, 1994, 1996). In these models, network behavior becomes rule-like with learning, but rules are not the actual representations of knowledge and rule firing is not the mechanism for cognitive processing. Rules are instead high-level, epi-phenomenal characterizations of what is happening at the sub-symbolic level of unit activations and connection weights. Among the many advantages of implementing rule-like behavior in neural activity are acquisition of non-normative rules, natural variation across problems and individuals, theoretical integration of perceptual and cognitive phenomena, and achievement of the right degree of crispness in knowledge representations. Several network predictions were confirmed in subsequent psychological studies.

An Encoder Version of Cascade-correlation

An apparent problem for using standard cascade-correlation in encoder problems is that it creates many cross-connections that bypass hidden units. The most troublesome of these for encoder simulations are the direct connections from input to output units, which could solve any encoder problem in a trivial way by rapidly learning weights of 1 between an input unit and its corresponding output unit. The solution is to freeze these direct input-to-output links to have values of 0, not modifiable by subsequent learning. As with back-propagation encoder networks, all of the computation must then employ hidden units.

Coding the Marcus et al. Experiments

The coding scheme for simulation of these experiments is a straightforward translation of words into an analog representation of real numbers. In such analog representations, degree of activation encodes distinct inputs and outputs. The assignment of words to numbers is arbitrary but consistent. In the training patterns, the four levels of A (*ga*, *li*, *ni*, *ta*) are represented by the numbers 1, 3, 5, and 7, respectively, and the four levels of B (*ti*, *na*, *gi*, and *la*) by the numbers 2, 4, 6, and 8, respectively. Hence, the ABA sentences *ga ti ga* and *li na li* are represented by 1 2 1 and 3 4 3, respectively. The ABB sentences *ga ti ti* and *ga na na* are represented by 1 2 2 and 1 4 4, respectively. The test patterns have values not used in training, but are interpolated within the training values: 2.5 for *wo*, 3.5 for *fe*, 5.5 for *de*, and 6.5 for *ko*. Thus, the ABA test sentence *wo fe wo* is represented by 2.5 3.5 2.5; and the ABB test sentence *de ko ko* is represented by 5.5 6.5 6.5.

In previous simulations with cascade-correlation networks, we have found that analog coding schemes often enable excellent learning and generalization. The use of analog representations is also supported by many psychological studies, particularly on numerical operations (e.g., Gelman & Gallistel, 1978, 1992).

Procedure

Sixteen cascade-correlation networks were run in each of the conditions of Marcus et al.'s Experiments 1 and 3. Each network, starting with its own randomly determined connection weights, including those initial weights used for candidate hidden units, corresponds to a unique infant. In each network, there were three input units to represent each of the three words in a sentence, and three output units to represent the target response, that is, the same three-word sentence. The output units had linear activation functions to enable their approximation of real numbers. The encoder option ensured that direct input-to-output connections were frozen at 0, so hidden units, with sigmoid activation functions would have to be recruited.

All cascade-correlation parameters were equal to Fahlman's default values with the following exceptions. Score-threshold, the tolerated difference between target and actual outputs was raised from the default of 0.4 to 1.0 in order to reduce the crispness of the rules learned by the network. Without this increased sloppiness, networks would never reverse the difference between consistent and inconsistent test sentences as did one of Marcus et al.'s infants. Training continued until all output units produced activations within score-threshold of their targets. There are also parameters for input-patience and output-patience with default settings of 8. They represent the number of epochs allowed to pass with little or no increase in correlation or

reduction in error, respectively, before shifting phase.[1] Cascade-correlation alternates between input and output phases, depending on whether a hidden unit is being recruited or weights going into output units are being adjusted, respectively. I changed these two patience values to 1, partly to increase sloppiness in network learning and partly because, on this problem, performance did not improve much after it failed to improve on a single epoch.

Results

Results in terms of error on the two types of test patterns are presented in Table 1 for the simulation of Experiment 1. Error on the test patterns was subjected to a repeated measures ANOVA in which condition (1 vs. 2) served as a between network factor and test pattern (consistent vs. inconsistent) served as a repeated measure. Neither the main effect of condition or the condition x test pattern interaction was significant. However, there was a substantial main effect of test pattern, $F(1,30) = 228$, $p < .0001$. As revealed in Table 1, there was more error to the inconsistent test patterns than to the consistent test patterns. With error considered to be equivalent to the need for further cognitive processing, this result mirrors that found with Marcus et al.'s infant participants. In a further parallel to the infant study, one network produced a reversal of the general trend, i.e., it showed (slightly) more error to the consistent test patterns than to the inconsistent test patterns.

Analogous results for the simulation of Experiment 3 are presented in Table 2 A similar ANOVA yielded only a substantial main effect of test pattern, $F(1,30) = 356$, $p < .0001$. Again, as revealed in Table 2, there was more error to inconsistent test patterns than to consistent test patterns. And again, there was one network with a reversal of the general trend, i.e., it showed (slightly) more error to the consistent test patterns than to the inconsistent test patterns.

Apart from needing a score-threshold of at least 1.0 to produce any reversals on the test patterns, other simulations showed that results were robust against systematic variation in the score-threshold and patience parameters.

A plot of results for one representative network is presented in Figure 1. It shows a negatively accelerated decrease in error over output epochs on the training patterns, much like the shape of declining attention in infant habituation experiments. After complete success with the training patterns, the consistent test patterns likewise show very little error, but the inconsistent test patterns show considerable error recovery, much like dishabituation of attention in infants. The epochs at which hidden units are recruited are marked with diamonds just above the training errors. As in other cascade-correlation simulations, it is noteworthy that error often decreases sharply after a new hidden unit is recruited.

Preliminary analyses of the knowledge representations learned by these networks suggest that the hidden units cluster on two fundamental components, each of which is sensitive to variation in both the A and B categories of words.[2] The two-dimensional nature of this problem was further verified by PCAs of the raw training data in each experimental condition.

Discussion

The simulation results show that a neural network model without variable-laden symbolic rules can indeed simulate the results of Marcus et al.'s (1999) infant habituation experiments. Like the infants, the networks showed gradual habituation to a repeated syntactic form, and recovery of interest to an inconsistent novel form but not to a consistent novel form. Even the occasional reversal preference by a single individual was captured. These results show that Marcus et al.'s findings with infants do not uniquely require a symbolic rule-based account. It may well turn out that some computation in humans is based on explicit symbolic rules, but the Marcus et al. data do not provide definitive proof for this claim as it applies to infants. Pinker's (1999) argument that the Marcus et al. data suggest an innate rule-learning capacity seems, at best, premature.

The key feature of the present simulations would appear to be the use of analog encoding for the input and output words composing a sentence. Generalization to novel items is known to be facilitated by analog coding schemes in which activation intensity corresponds to particular representations (Jackson, 1997).[3] Although the details of Marcus et al.'s (1999) unsuccessful simulations were not published, it might be speculated that they employed non-analog binary codes. The reason such codes might not generalize is that novel items are coded on units with untrained connection weights. Such failures are analogous to expecting that I can speak Spanish just because someone I never interact with has learned to do so.

The use of analog coding is not merely a way of smuggling in variable binding. Analog coding by itself does not implement variable binding because assignments of values to input units are lost as activation is propagated forward onto non-linear hidden units. In explicit variable-binding schemes, assignments of values to variables are preserved for use in later computation.

It is possible that other neural network algorithms, such as back-propagation or auto-association, would be able to simulate these habituation data, using analog or some other coding scheme. If so, differences between successful algorithms could be explored for relative accuracy in accounting for the psychological data.

Other future work on this model might profitably address the issue of whether a successful neural network model could use a more realistic phonetic coding of the input. Very few neural network models attempt to cover everything from raw stimuli to high-level cognitive manipulations. But

[1] An epoch is a presentation of all of the training patterns.

[2] These analyses are based on PCA of network contributions (Shultz, Oshima-Takane, & Takane, 1995).

[3] The present networks generalize to A and B syntactic categories even outside of the range of the training patterns.

such extensions would generate more complete understanding of psychological phenomena. The finding that phonemes vary continuously on sonority suggests that a more realistic analog encoding might be feasible (Vroomen, van den Bosch, & de Gelder, 1998). Research on such realistic coding schemes would be necessary to simulate Marcus et al.'s Experiment 2, which was designed with particular phonetic properties in mind.

Another useful extension might involve the use of recurrent networks. Non-recurrent networks and recurrent networks can both learn temporal problems. The essential difference between them is a trading of space for time in the input coding (Hertz et al., 1991). Recurrent networks process inputs in sequence over time, allowing for sequences of indeterminate length; non-recurrent networks represent inputs on different input units simultaneously. There is a recurrent version of cascade-correlation that might be interesting to try on these sentence habituation problems.

It is worth stressing that the present model is not the definitive treatment of habituation. The process of habituation is still poorly understood and there are many phenomena in the habituation literature that would need to be accounted for by any comprehensive model. This study is essentially a demonstration that the Marcus et al. (1999) data can be covered by a neural network model.

Because a turnabout is often considered fair play, I would like to issue a reciprocal challenge to those favoring symbolic rule-based models of human cognition to implement serious models of habituation phenomena. The habituation literature is extensive and contains some of the most important discoveries in developmental psychology. They are largely untouched by computational modeling. It does not suffice to merely re-describe psychological phenomena in terms of a few symbolic rules. It is critically important to implement working models that include not only knowledge representation and processing but also learning and development as appropriate. Among the significant challenges for rule-based models of habituation are clear links with new or standard theories of the habituation process, the gradual negatively accelerated shape of habituation curves, individual differences in habituation rates and occasional reversals of general trends, and the dishabituation differences reported by Marcus et al. (1999). One of the quickest cures for incorrect or inappropriate theoretical statements and models is the discipline of actually trying to implement a model that covers phenomena in a principled way.

Another set of infant habituation data that has been successfully simulated by an encoder neural network model concerns asymmetric exclusivity effects in infant memory and categorization (Mareschal & French, 1997). Infants learn the categories of *dog* and *cat*, but with some interesting asymmetries (Quinn, Eimas, & Rosenkrantz, 1993). Essentially, dogs are not included in the category of cats, but cats are included in the category of dogs.

At this point, the only successful models of infant habituation employ feed-forward connectionist models without explicit variable binding.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Comments on an earlier draft by Alan Bale, Dave Buckingham, Yasser Hashmi, Yuriko Oshima-Takane, Sylvain Sirois, and Yoshio Takane were gratefully received.

References

Buckingham, D., & Shultz, T. R. (1994). A connectionist model of the development of velocity, time, and distance concepts. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 72-77). Hillsdale, NJ: Lawrence Erlbaum.

Buckingham, D., & Shultz, T. R. (1996). Computational power and realistic cognitive development. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 507-511). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19*, 157-180.

Cohen, L. B. (1979). Our developing knowledge of infant perception and cognition. *American Psychologist, 34*, 894-899.

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

Fahlman, S. E., & Lebiere, C. (1990). The Cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2.* Los Altos, CA: Morgan Kaufmann.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gelman, R., & Gallistel, C. R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*, 43-74.

Haith, M. M. (1990). Progress in the understanding of sensory and perceptual processes in early infancy. *Merrill-Palmer Quarterly, 36*, 1-26.

Hashmi, Y., & Shultz, T. R. (1998). A neural network model of number comparison. (Submitted for publication).

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison Wesley.

Jackson, T. O. (1997). Data input and output representations. In E. Fiesler & R. Beale (Eds.), *Handbook of neural computation*. Oxford: Oxford University Press.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80.

Mareschal, D. & French, R. M. (1997). A connectionist account of interference effects in early infant memory and

categorization. In *Proceedings of the 19th annual conference of the Cognitive Science Society* (pp. 484-489). Mahwah, NJ: LEA.

Mareschal, D., & Shultz, T. R. (in press). Development of children's seriation: A connectionist approach. *Connection Science*.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5*, 193-207.

Pinker, S. (1999). Out of the minds of babes. *Science, 283*, 40-41.

Quartz, S. R, & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences, 20*, 537-596.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. *Advances in Infancy Research, 10*, 1-36.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception, 22*, 463-475.

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science, 1*, 103-126.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro, & T. K. Leen, (Eds). *Advances in Neural Information Processing Systems 7* (pp. 601-608). Cambridge, MA: MIT Press.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.

Slater, A. (1995). Visual perception and memory at birth. *Advances in Infancy Research, 9*, 107-125.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Hillsdale, NJ: Erlbaum.

Takane, Y., Oshima-Takane, Y., & Shultz, T. R. (1995). Network analyses: The case of first and second person pronouns. *Proceedings of the 1995 IEEE International Conference on Systems, Man and Cybernetics* (pp. 3594-3599).

Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes, 13*, 193-220.

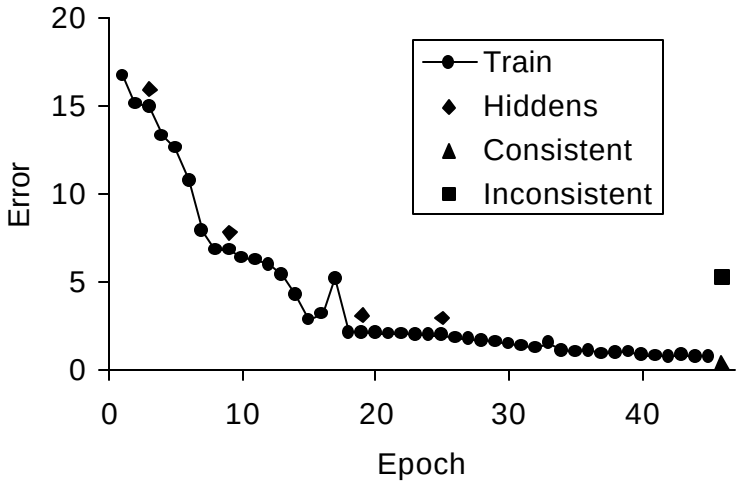

Figure 1: Results for a representative network in the ABA condition of the simulation of Experiment 1. Error on the training patterns decreases in a negatively accelerated fashion over time, representing habituation. Error remains low for the consistent test, but increases for the inconsistent test, demonstrating dishabituation to novelty. The diamond shapes represent the epochs at which hidden units were recruited.[4]

[4] Error is divided by the number of patterns and plotted over output-phase epochs. Input-phase epochs are not included in such plots because there is no change in network performance during input phases. The first three output epochs are omitted from this plot to improve clarity because error starts quite high and drops dramatically by virtue of adjustment of weights from the bias unit. The bias unit is always on, with an activation of 1, regardless of the input pattern. It has trainable connection weights to all non-input units, specifying their resting levels of activation.