Developmental Science 3:4 (2000), pp 442–456

PAPER

# Artificial grammar learning by infants: an auto-associator perspective

Sylvain Sirois, David Buckingham and Thomas R. Shultz

*Laboratory for Natural and Simulated Cognition, McGill University, Canada*

## Abstract

*This paper reviews a recent article suggesting that infants use a system of algebraic rules to learn an artificial grammar (Marcus, Vijayan, Bandi Rao & Vishton, Rule learning by seven-month-old infants.* Science, **183** (1999), 77–80). *In three reported experiments, infants exhibited increased responding to auditory strings that violated the pattern of elements they were habituated to. We argue that a perceptual interpretation is more parsimonious, as well as more consistent with a broad array of habituation data, and we report successful neural network simulations that implement this lower-level interpretation. In the discussion, we discuss how our model relates to other habituation research, and how it compares to other neural network models of habituation in general, and models of the Marcus* et al. *(1999) task specifically.*

Over the last 30 years the habituation paradigm has become a widely used technique for investigating infant cognition. Developmental psychologists initially used the habituation paradigm to study sensory and perceptual processes in infants in much the same way as it had been used in animal research (Clifton & Nelson, 1976; Malcuit, Pomerleau & Lamarre, 1988). That is, researchers focused on 'the simplest form of learning' (Thorpe, 1963). More recently, however, the habituation paradigm has been used to suggest a broad range of early conceptual abilities in infants.

In this paper, our focus is on a recent habituation study suggesting that 7-month-old infants are capable of grammar learning by means of a system of algebraic rules (Marcus, Vijayan, Bandi Rao & Vishton, 1999). As an alternative explanation, we propose a modality-independent, feature-independent model of habituation that implements lower-level processing. We focus on the Marcus *et al.* (1999) report for two reasons. First, given the landmark importance this study could have (Pinker, 1999), it is worthy of further attention. Second, it reports that a rule-based account is to be preferred, because one class of neural networks fails to capture the data. We report neural network simulations that do capture the data.

The paper is organized as follows. The first section briefly discusses habituation as a tool for the investigation of high-level cognition. The procedure used by Marcus and his colleagues (1999) is presented, their interpretation is reviewed, and a simpler interpretation of their data is offered. We then present the auto-associator neural network as a potential model of infant habituation, and report on simulations using such networks that capture the infant data reported by Marcus *et al.* (1999). In a general discussion, we present our results in the context of alternative interpretations and models.

## Habituation: a brief introduction

The habituation paradigm was originally devised to investigate the sensory and perceptual abilities of organisms (Clifton & Nelson, 1976; Malcuit *et al.*, 1988; Haith, 1998) by repeatedly presenting a stimulus (or set of stimuli) until there was a decrease of behavioral response (such as heart-rate changes, galvanic skin response changes, head turns, and/or visual fixation). However, Sokolov's theoretical interpretation

Address for correspondence: Sylvain Sirois, Department of Psychology, McGill University, 1205 Dr Penfield Avenue, Montréal, Québec, Canada H3A 1B1; e-mail: sirois@psych.mcgill.ca

of the decrease in responding provided developmental psychologists with a unique tool to investigate preverbal cognitive abilities (Clifton & Nelson, 1976; Malcuit *et al.*, 1988). Sokolov (1963) argued that, during habituation, organisms build an internal schema of the stimuli. Over trials, the discrepancy between this schema and input decreases through learning, and thus the organism stops responding to insignificant differences. A novel stimulus, however, could produce renewed responding (i.e. recovery) when there is perceived discrepancy between that stimulus event and the schema.[1]

A variation of Sokolov's representational account of habituation is often used to interpret the results of habituation experiments designed to investigate cognitive abilities in infants (e.g. Baillargeon, 1987; Spelke, Breinlinger, Macomber & Jacobson, 1992; Wynn, 1992, 1995; Marcus *et al.*, 1999). The basic tenet is that if infants have particular conceptual knowledge they will appreciate violations of this knowledge by attending longer to unusual, inconsistent or impossible events (Spelke, 1998). Therefore, in the test phase of habituation experiments, researchers can measure whether infants exhibit differential attention to test events violating the knowledge under investigation, compared with equally novel and perceptually similar test events that are consistent with this knowledge.

Using this approach, researchers have suggested that infants have conceptual knowledge about object permanence (Baillargeon, Spelke & Wasserman, 1985; Baillargeon, 1987), object properties such as solidity and continuity (Spelke *et al.*, 1992) and integer numbers (Starkey, Spelke & Gelman, 1990; Wynn, 1992, 1995). However, perceptual-level interpretations of the tasks used to infer such conceptual knowledge have also been proposed (Mareschal, Plunkett & Harris, 1995; Bogartz *et al.*, 1997; Mix, Levine & Huttenlocher, 1997; Munakata, McClelland, Johnson & Siegler, 1997; Cohen, 1998; Haith, 1998; Sirois *et al.*, in preparation). Furthermore, on the grounds of parsimony, Haith (1998) argues that perceptual-level accounts based on principles such as novelty, familiarity, salience and discrepancy should be favored over conceptual-level accounts if the former cannot be ruled out.

One question then is whether a perceptual-level account of 'rule learning in 7-month-old infants' (Marcus *et al.*, 1999) can be sustained. In the next section, we review their study.

## Abstract algebraic rules in infants reviewed

In Marcus *et al.*'s (1999) first experiment, infants were habituated to 16 three-syllable sentences that followed either an 'ABA' or an 'ABB' grammar. Examples of 'ABA' sentences are 'ga ti ga' and 'li na li', and examples of 'ABB' sentences are 'ga ti ti' and 'li na na'. The 16 sentences were constructed by using four A syllables and four B syllables, resulting in 16 combinations for each grammar. A 250 ms pause was placed between syllables within a sentence, and sentences were separated by 1 s pauses.

In typical habituation experiments, an individual stimulus event is presented repeatedly (or continuously, depending on the nature of the stimuli) until the infant's attention has not been directed at the event for a specific amount of time, or until the trial has exceeded some duration criterion. The next stimulus event can then be presented. Marcus and colleagues (1999), however, used a procedure adapted from Saffran, Aslin and Newport (1996) for the habituation (or familiarization) phase. Their 16 sentences were used to create three consecutive blocks of 16 habituation trials. Within a block, each sentence appeared once, in a random position, and was not repeatedly presented. There was no pause between blocks except for the 1 s interval between all sentences. As such, infants in this experiment heard a continuous speech stream like 'ga ti ga | ga la ga | li na li | ta ti ta | ...' (where spaces identify 250 ms pauses and ' | ' is the 1 s pause between sentences) until the third block had been completed. These sentences were played from both left and right speakers simultaneously.

During the habituation phase, a yellow light flashed in front of the infants in order to draw their attention. When the number of prescribed trials was reached, habituation was assumed and testing began.[2] Two 'ABA' and two 'ABB' sentences, each constructed with novel syllables (i.e. syllables not used in the habituation phase), served as test sentences. As such, both types of test sentences were equally novel.

[1] Current interpretations of habituation such as information processing approaches (e.g. Zelazo, 1988; Cohen, 1998) and perceptual processing approaches (e.g. Bogartz, Shinskey & Speaker, 1997; Haith, 1998) are similar to Sokolov's (1963) representational account; they differ mainly in their emphasis on specific processes (Sirois, Debbané & Zelazo, in preparation).

[2] Familiarization procedures run the risk of providing infants with insufficient training, or increasing variability in test results by leaving infants at various stages of encoding at the end of the habituation phase (Clifton & Nelson, 1976; Malcuit *et al.*, 1988). This concern may not apply to Marcus *et al.* (1999), because they report the typical novelty preference in the test phase of their experiments. Their results would probably replicate in variants that actually measure habituation and end training when habituation has been observed (Clifton & Nelson, 1976).

At the onset of a test trial, the central light flashed to draw the infant's attention. When the infant fixated on this light, it was extinguished and one of the red side lights began flashing. The test sentence would begin playing from the corresponding left or right speaker when the infant had turned toward the flashing light. This test sentence was presented repeatedly until the infant looked away from the flashing light, or until 15 s had elapsed. Infants were presented with three blocks of test sentences, with all four test sentences in random order within each block. Because infants were habituated to either 'ABA' or 'ABB' sentences, half of the test sentences were consistent and the other half were inconsistent with the training grammar.

The results showed that looking times of infants were longer for inconsistent sentences. The authors pointed out that an interpretation based on transitional probabilities (e.g. Saffran *et al.*, 1996) cannot account for these data. Transitional probabilities (i.e. the probability that a given syllable follows another one) could certainly be learned in the habituation phase but would be of no use when novel test syllables are introduced. For the same reason, a system that notes discrepancies with stored sequences of words cannot account for differential attention to consistent and inconsistent sentences in the test phase (Marcus *et al.*, 1999). The authors argued that abstract algebraic rules that represent relationships between variables (e.g. 'item $x$ is the same as item $y$') could account for the data.

In their second experiment, the authors controlled for some overlap in sequences of phonetic features between training and testing sets that may have been a confound. Some of the training sentences in the 'ABA' habituation set, for example, had a voiced–unvoiced–voiced sequence of consonants. Each 'ABA' test sentence had the same sequence, whereas each 'ABB' sentence had a voiced–unvoiced–unvoiced sequence. In order to control for the possibility that infants relied on phonetic features instead of deriving an abstract rule,[3] the authors replicated the first experiment with a phonetic control. That is, all syllables used in the habituation set were voiced, whereas they varied as before in both test sets. The results mirrored data obtained in the first experiment: infants attended longer to grammatically inconsistent sentences.

In a third experiment, the authors controlled for another possible confound. In 'ABB' sentences, there is an immediate duplication of syllables that is not found in 'ABA' sentences. Infants could thus have distinguished the two grammars on the presence or absence of such duplication. Using the same habituation-recovery method from Experiments 1 and 2, Marcus and his colleagues (1999) habituated infants with either 'AAB' or 'ABB' sentences and used novel sentences from both grammars in the test phase. As in the first two experiments, infants attended longer to grammatically inconsistent test sentences.

Marcus and his colleagues (1999) concluded from these three experiments that infants must have the ability to extract abstract algebraic rules in order to react to inconsistent sentences. As further evidence for the rule-based interpretation, Marcus *et al.* (1999) discuss neural network simulations they conducted, which failed to capture the data of their experiments.

In his companion article, Pinker (1999) argues that the report from Marcus and colleagues 'suggests that one of the mechanisms that makes computers intelligent – manipulating symbols according to rules – may be a basic mechanism of the human brain as well' (p. 40). Pinker's (1999) contention is that Marcus and colleagues have sustained a central claim in classic psycho-linguistic theory: infants are innately endowed with symbol-manipulating machinery that enables them to acquire language. This conclusion assumes that Marcus *et al.*'s (1999) interpretation is correct. A case can be made, however, that it is premature.

Marcus and colleagues (1999) overstate the inability of statistical learning mechanisms to capture the data (McClelland & Plaut, 1999). There are inherent statistical regularities in Marcus *et al.*'s (1999) stimuli. By introducing novel words in the testing phase, the authors did not simultaneously remove the covariance structure in each sentence, which is identical for consistent test sentences and habituation sentences. This covariance structure is perceptual in nature: the syllables are organized in such a way that two of them are always redundant.

Marcus and colleagues (1999) use labels such as 'ABA' to describe the grammar underlying the stimuli they present to infants. Whereas an 'ABA' grammar was a methodological given for the authors, it would necessarily be an induction for infants.[4] This induction can only be assumed from differential looking times in the testing phase; it is not an empirical fact and, as such,

[3] We do not understand how learning about phonetic features cannot qualify as algebraic rules. From Marcus and colleagues' (1999) example of such rules, 'item $x$ is the same as item $y$', nothing would seem to prevent the operator 'is the same' from applying to phonetic features of $x$ and $y$.

[4] As Kemler (1981) noted, what is actually perceived as meaningful by infants in habituation experiments cannot be assumed to correspond even to adults' simplest formal description of the stimuli. McClelland and Plaut (1999) make a similar point concerning the stimuli used by Marcus and colleagues (1999).

alternative interpretations are equally valid. McClelland and Plaut (1999) discussed several such alternatives. We actually favor an alternative that they did not discuss.

This alternative is that infants formed a prototypical pattern of the perceptual covariance structure in the habituation corpus, and that a mismatch between this prototype and inconsistent test items generated longer looking times (e.g. Sokolov, 1963; Younger & Cohen, 1985; Cohen, 1998). We favor this other alternative because it is consistent with a broad array of habituation data and it has already been proposed in the habituation literature. A label such as 'ABA', which Marcus *et al.* (1999) refer to as a grammar, can actually be construed as a perceptual-level prototypical pattern. Because this prototype would be an abstraction of the habituation corpus, it is conceivable that it would match consistent test items, which have the same perceptual structure, albeit with new elements.

The distinction between this alternative interpretation and abstract algebraic rules is not merely one of terminology. In Marcus *et al.*'s (1999) interpretation, perceptual input would be transformed into an appropriate symbolic format such that it can serve as input for higher-level rule-based computations using variables. In the prototype alternative, computations are performed on the perceptual input directly and do not require subsequent, higher-order machinery.

We suggest that the auto-associator neural network is a likely candidate to model infant habituation, and that it would implement this alternative, lower-level interpretation. Moreover, it can model the temporal nature of looking times in habituation experiments. In order to substantiate these claims, the next section presents the relevant properties of this neural network architecture, as well as a series of successful simulations of Marcus *et al.*'s (1999) experiments.

## An auto-associator model of habituation

The auto-associator (Anderson, Silverstein, Ritz & Jones, 1977; Kohonen, 1977), as depicted in Figure 1, consists of a set of simple processing units, fully interconnected with one another. This square matrix of connections, called weights, processes the internal activity in the network.[5] When a pattern of external

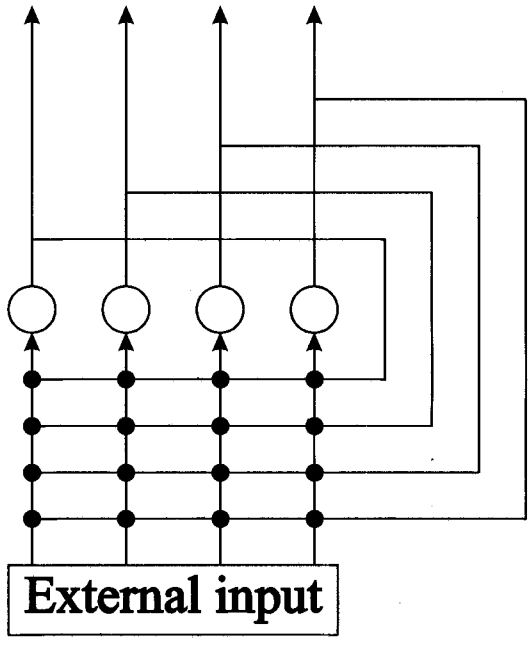


**Figure 1** *A generic auto-associator network. External input is first presented to the network, activating the corresponding units (large white circles). Then, each unit sends its activation via weighted connections (lines) to all units, including itself. This process can be repeated any number of cycles. Connections leading away from the internal units would enable the network to pass its activations to other networks.*

input is presented to the network, the corresponding units are activated, and these propagate their activations to all units in the network, including themselves. This internal circulation of activations can take place for several cycles. Weights and activations are represented as real numbers. Units stimulate or inhibit one another on each processing cycle. Because weights can be modified as a function of experience, networks can learn regularities between features across a set of training patterns.

A variety of implementations of the auto-associator have been proposed (Anderson, 1977; Kohonen, 1977; McClelland & Rumelhart, 1985). Some models use a linear activation function, in which the activation of a given unit is equal to the input it receives. The problem with this function is that, over cycles, activations do not settle but continue to grow unbounded. Alternative activation functions involve some form of clipping of the linear function; that is, activation values are equal to the input but whatever exceeds a lower and upper bound is clipped. Other models use nonlinear activation functions, such as the sigmoid function (discussed later). Models also differ with respect to the learning rule used to update weights. Some models prevent self-connec-

[5] Although we discuss the auto-associator as a simple neural network, it is worth noting that it is actually a general class of networks, and that many other types of network architectures are restricted implementations of the auto-associator. By restricting the connectivity and input of the auto-associator, one can implement a multilayered feedforward network, for example.

tions (i.e. the connection between a unit and itself), such that units learn only their relationship to the other units in the network. Despite these variations, the auto-associator excels at capturing covariation in training sets. Auto-associator networks are particularly suited for pattern completion tasks. Given incomplete input, they will generate a complete pattern based on what they were trained on.

In order to model habituation experiments, we devised our own implementation of the auto-associator. For unit activation, we used the sigmoid function. This nonlinear function keeps activation values within the −0.5 to 0.5 range, thus providing natural clipping. The sigmoid activation of a unit is computed as

$$a_i = \frac{1}{1 + \exp(-\lambda \text{ net}_i)} - 0.5 \qquad (1)$$

where $a_i$ is the activation of unit $i$, exp is the natural log, $\lambda$ is a constant called the temperature parameter, and $\text{net}_i$ is the input that unit $i$ receives. The output of this function is shown in Figure 2. As the figure implies, the activation of sigmoid units is constrained between −0.5 and 0.5. Unit behavior shows more variation to small changes in input when this input is close to zero than to the same changes when absolute input is large. The temperature parameter affects the slope of the function, but not its range. Values above 1 result in a steeper slope, whereas values between 0 and 1 result in a flatter slope.

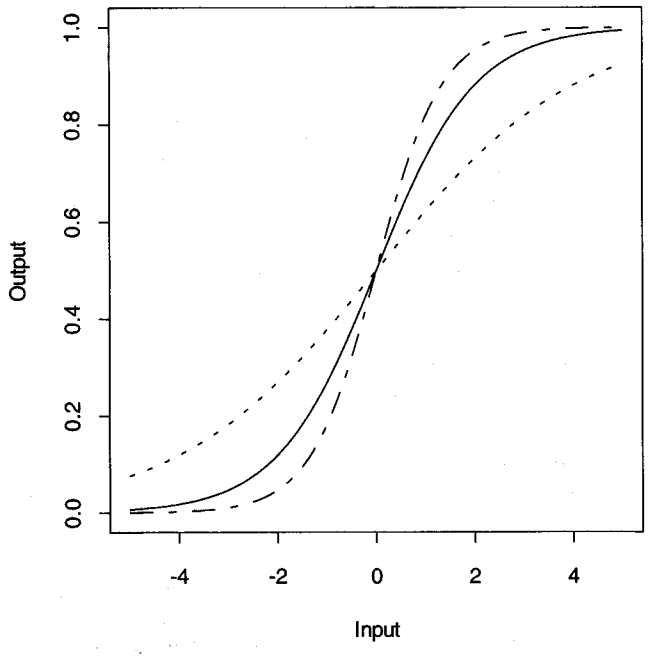


**Figure 2** *Sigmoid activation function. Curves are shown for temperature values of 0.5 (dotted line), 1 (solid line) and 1.5 (broken line).*

In the auto-associator, the input to a unit is a function of both internal activity and external input, and is computed by

$$\text{net}_i = \sum_j w_{ij} a_j + \eta E_i - \xi a_i \qquad (2)$$

where $\text{net}_i$ is the input unit $i$ receives, $j$ is the index of the sending unit, $w_{ij}$ is the weighted connection between unit $i$ and unit $j$, $a_j$ is the activation of the sending unit $j$, $\eta$ is a constant we refer to as the input scalar, $E_i$ is the external input of unit $i$, and $\xi$ is the decay parameter. For a given unit, then, the input is the sum of weighted activations from all units (including itself), plus some proportion of the external signal, minus some proportion of its previous activation. The input scalar $\eta$ allows the external signal to be amplified ($\eta > 1$) or weakened ($0 < \eta < 1$).

In our implementation of the auto-associator, input processing is a four-step process. First, the network is presented with an external input pattern and the net input to units is computed. There is initially no internal activation, so the net input computed with equation (2) on this first step is only the external input (i.e. the first and third terms on the right-hand side of equation (2) equal 0 because $a_j$ is initially 0). Second, by applying equation (1) to the net input, we obtain the unit's activation corresponding to the external pattern. Third, these activations are used in equation (2) to compute an updated net input of the network based on both external input and internal activity. Lastly, the net input from the third step is used to compute the updated unit activation values with equation (1). These values from the fourth step represent the network's response to external information combined with its internal activity.[6]

After the fourth step, when unit activations reflect internal representations based on external information, weights in the network are modified to implement learning. The learning rule used in our auto-associator networks is the delta rule, which computes weight changes as

$$\Delta w_{ij} = \lambda(E_i - a_i)a_j \qquad (3)$$

where $\Delta w_{ij}$ is the amount by which the weight between sending unit $j$ and receiving unit $i$ is to be changed, $\lambda$ is the learning rate constant, $E_i$ is the external input for unit $i$, $a_i$ is the activation of unit $i$, and $a_j$ is the activation of the

[6] In some implementations of the auto-associator, the third and fourth steps are repeated for a number of times, a process called internal cycling. We did not implement this additional procedure to model Marcus *et al.*'s (1999) experiments. This is discussed in more detail later.

sending unit *j*. This learning rule implies that connection weights between units are changed as a function of the discrepancy between the activation of a unit and the external input it receives, and of the activation of the sending unit. Over learning steps, the network is therefore trained to reproduce the external input by allowing larger weights between units that have correlated activations across the training set. The higher the correlation, the larger the weight. And as activation values get closer to the external input, weight changes decrease. This naturally prevents weights from growing too large.

We argue that such networks can implement the habituation-recovery paradigm in a way that accurately maps the experimental procedure. Networks can be trained on habituation patterns until all activations on all patterns change by less than some threshold value between epochs (an epoch is a presentation of all training patterns). Thus, we can stop the habituation phase when the network has achieved a stable representation of the habituation set through learning. This can be assumed to underlie the decrease of responding in infants, as further processing would pointlessly tap attentional resources.

When habituation is observed in networks, they can then be presented with individual testing patterns, and the number of presentations required to learn these test items can be taken as an index of processing time. It is therefore possible to test networks for differential processing on novel consistent and inconsistent items in a way that is analogous to the empirical procedure. Both the dependent measure used in our simulations (i.e. the number of processing steps) and the attentional measures used with infants (e.g. looking times) are temporal in nature.

In order to model Marcus and colleagues' (1999) experiments, we used an arbitrary coding scheme to represent the syllables that formed sentences. Sixteen syllables were created by using all possible combinations of four binary values, as depicted in Table 1. Within this encoding framework, a given syllable is coded on four units, and networks that process three-syllable sentences use 12 units. The use of this arbitrary encoding scheme makes our simulations a general account of the regularities underlying the Marcus *et al.* (1999) data, as the model is not rooted in the specifics of these experiments. The questions we ask of the model are can the auto-associator (a) capture the structure of, for example, 'ABA' habituation events and (b) show differential recovery to consistent and inconsistent test events? A successful answer to both questions within this arbitrary scheme would suggest a modality-independent, feature-independent model of habituation.

We now report of a series of simulations of the Marcus *et al.* (1999) habituation experiments. Simulation 1 investigated whether our auto-associator model could learn an 'ABA'-type structure and generalize to novel sentences by exhibiting differential processing of consistent and inconsistent stimuli. The arbitrary coding scheme we used does not depend on specific phonetic features. As such, this simulation models Experiment 1 in Marcus *et al.* (1999). Simulation 2 presents networks similarly trained and tested on 'AAB' and 'ABB' structures, modeling the third experiment of Marcus and colleagues (1999).

**Table 1** *Distributed binary encoding scheme for 16 syllables*

| Pattern | Unit 1 | Unit 2 | Unit 3 | Unit 4 |
|---|---|---|---|---|
| 1 | −0.5 | −0.5 | −0.5 | −0.5 |
| 2 | −0.5 | −0.5 | −0.5 | 0.5 |
| 3 | −0.5 | −0.5 | 0.5 | −0.5 |
| 4 | −0.5 | 0.5 | −0.5 | −0.5 |
| 5 | 0.5 | −0.5 | −0.5 | −0.5 |
| 6 | −0.5 | −0.5 | 0.5 | 0.5 |
| 7 | −0.5 | 0.5 | −0.5 | 0.5 |
| 8 | 0.5 | −0.5 | −0.5 | 0.5 |
| 9 | −0.5 | 0.5 | 0.5 | −0.5 |
| 10 | 0.5 | −0.5 | 0.5 | −0.5 |
| 11 | 0.5 | 0.5 | −0.5 | −0.5 |
| 12 | −0.5 | 0.5 | 0.5 | 0.5 |
| 13 | 0.5 | −0.5 | 0.5 | 0.5 |
| 14 | 0.5 | 0.5 | −0.5 | 0.5 |
| 15 | 0.5 | 0.5 | 0.5 | −0.5 |
| 16 | 0.5 | 0.5 | 0.5 | 0.5 |

### *Simulation 1: ABA versus ABB*

This simulation models Experiment 1 from Marcus *et al.* (1999). Networks were habituated on 'ABA' stimuli and tested on 'ABA' and 'ABB' stimuli consisting of novel syllables. Figure 3 outlines the procedure we used to implement sequential input. At time 1, the first syllable is presented to the network and the four-step procedure is used to compute activations. The net input for each unit is computed and then transformed by the activation function. These activations are circulated through the network and used to compute an updated net input, consisting of external information and internal activity. Activations are then updated, at which point weights are changed according to the learning rule. At time 2, the second syllable is introduced. The first syllable is still part of the external input but is decayed by some proportion of itself.[7] The four-step procedure to compute activations is applied, and weights are then

[7] This fading encoding representation scheme is consistent with a perceptual perspective on habituation which considers lingering sensory information (Haith, 1998; Sirois *et al.*, in preparation). See Ungerleider (1995) for a review of neurological support for this short-term perceptual effect.

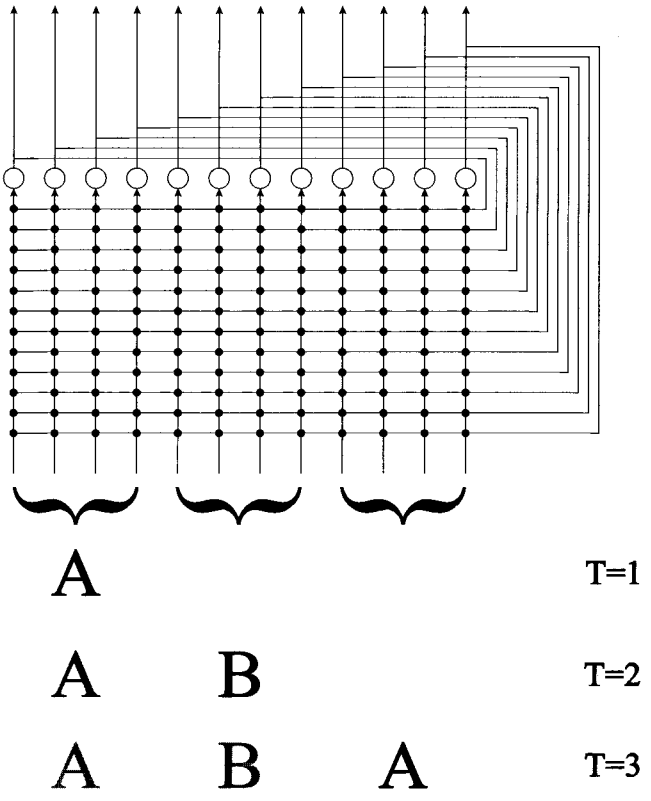


**Figure 3** *Depiction of the sequential auto-associator network. Three banks of units are used to process the three syllables in 'ABA' sentences. Each syllable is encoded on four binary units. The syllables are presented over three time steps: A, then B, and finally A. Previously presented syllables decay gradually as new syllables are presented (represented graphically by the fading letters).*

updated. During the first step, however, internal activations obtained from the syllable presented at time 1 are part of the net input. That is, when the second syllable is introduced, the network is already active from the first syllable. At time 3, the third syllable is introduced as external input, whereas the first two are decayed (resulting in the most decay for the first syllable). Activations are computed using the four-step procedure, and weights are further updated.

This sequential procedure is used for both habituation and testing patterns. After habituation, the networks are expected to require significantly more processing time to learn inconsistent test patterns compared with consistent ones. For both types of test patterns, learning time should decrease over blocks of test trials as networks adapt to the novel items through weight changes.

## Method

Thirty networks were used in this simulation. For each network, a unique set of training and testing patterns was created. From the bank of 16 syllables listed in Table 1, four were randomly selected (without replacement) as A and four as B to construct the habituation patterns. Combining all A and B syllables resulted in 16 'ABA' patterns. From the remaining eight syllables in the bank, four were randomly selected to create the testing patterns. These four syllables were randomly assigned in two pairs in order to form two sets of consistent and inconsistent test patterns (such as 'wo fe fe', 'wo fe wo', and 'la ti ti', 'la ti la'), which is how testing items were devised in Marcus *et al.* (1999).

Networks were habituated to 'ABA' patterns, and tested on 'ABA' and 'ABB' patterns. In the habituation phase, networks were trained on all 16 habituation patterns, which were presented in random order in each epoch. As in Marcus *et al.* (1999), each sentence was presented once per block of 16 trials. When the last syllable of a sentence had been presented and weights had been updated, activations were reset to zero and the next sentence was presented. Training continued until the change in activation on all units and on all patterns between the current and previous epoch was below 0.005, at which point the network was considered to have habituated to the training set. If a network reached a 50 epoch limit before activation changes were below criteria, training would be stopped and testing would begin as for networks that had habituated according to our criteria.

Following Marcus *et al.*'s (1999) procedure, there were three consecutive blocks of testing in our simulations. In each block, the four testing patterns were presented in random order to the network. For each individual test pattern, activations were repeatedly computed (and the weights updated) until the change in activation for all units between the current and previous presentation was below 0.005. At that point, the test pattern was considered to no longer require further processing. We refer to the repeated presentation of individual test sentences as testing cycles, and these will be the dependent variable.

It is worth pointing out that our networks always process information the same way, whether they are in the habituation phase or the test phase. A pattern is presented sequentially and, as each syllable is presented, the four-step procedure is used to compute activations and weights are then updated. What distinguishes habituation and testing is whether an individual pattern is presented repeatedly, but this is a constraint from Marcus *et al.*'s (1999) procedure that affects *what* networks process, but not *how* they process it.

The parameter values used in this simulation are 0.04 for the learning rate, 1 for the temperature of the sigmoid function, 0.005 for the decay parameter, and 3

for the input scalar. For each network, weights are initially set to zero. The parameter value used to decay previous syllables was 0.9, which implies that already present syllables were decayed by 10% at each of times 2 and 3 in a 'sentence' presentation.

## Results

Networks in this simulation required an average of 24.1 epochs to learn to criterion in the habituation phase (SD = 1.92). All 30 networks habituated within the 50 epochs limit. Figure 4 plots the number of testing cycles required to learn each type of test pattern over testing blocks. The average number of testing cycles for consistent items was 3.4, 3.2 and 3.1 for test blocks 1, 2 and 3. For inconsistent items, the average number of testing cycles was 3.6, 3.4 and 3.2 over test blocks. Six of the networks required more testing cycles for consistent test items, on average, than for inconsistent items. Testing cycles required to learn were analyzed with a 2 by 3 by 2 repeated-measures analysis of variance, with type (consistent and inconsistent), block (1–3) and pattern (first and second) as within-subject factors. The analysis revealed a significant effect of type ($F(1, 29) = 5.69$, $p < 0.05$) as well as a significant effect of block ($F(2, 58) = 35.32$, $p < 0.05$). There was no significant effect of pattern, nor any significant interaction effect. Subsequent tests of within-subject contrasts revealed a significant linear trend for block ($F(1, 29) = 49.00$, $p < 0.05$).

Figure 5 depicts the connectivity of a representative network, i.e. a network that required 24 epochs of training during the habituation phase (the average from this simulation) and fewer pattern repetitions for consistent test items. The raw weight values shown in the figure were recorded after the habituation phase, before testing began. The diagram can be mapped onto the network representation in Figure 3, for clarity. Columns represent incoming weights for individual units, and rows represent outgoing weights from individual units. In this figure, larger boxes represent larger weights. Positive connections are represented with white boxes, and negative connections with black boxes. Weights with near-zero values do not appear in the figure. The weights along the diagonal represent self-connections. This network was habituated to an 'ABA' structure; redundant units were thus 1–4 and 9–12.

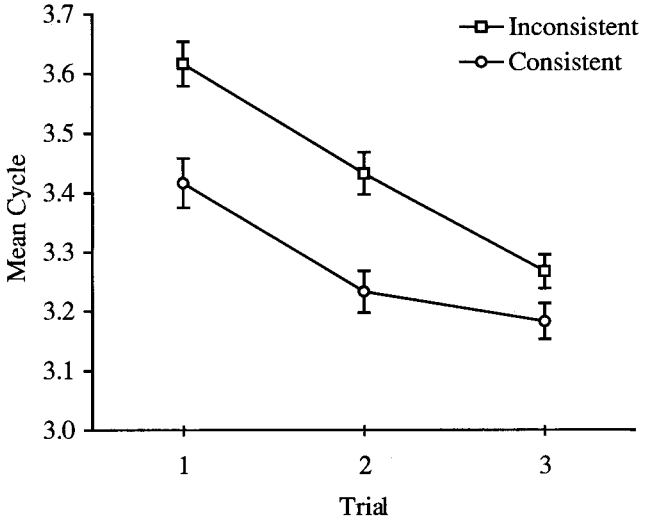


**Figure 4** *Mean number of testing cycles required for network learning of novel consistent (ABA) and inconsistent (ABB) test patterns by trial in Simulation 1. Error bars show standard deviations. (The range of the y axis in this figure and in Figure 7 has been restricted to the range of the data for clarity.)*

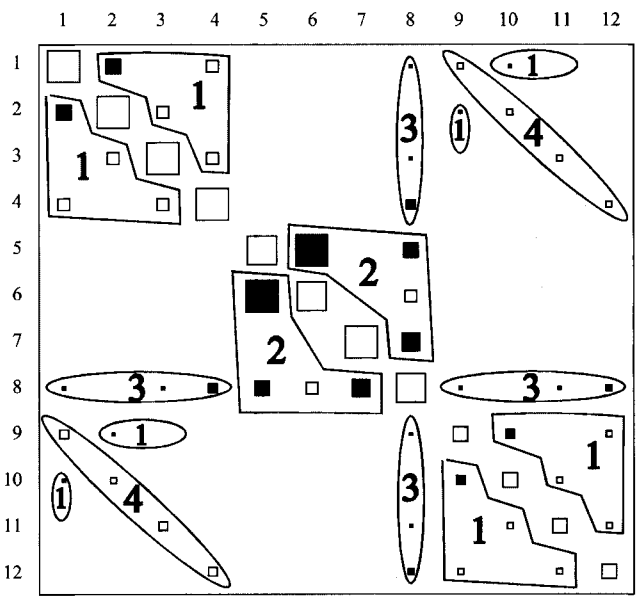


**Figure 5** *Weight diagram of a representative network, after habituation and before testing. Columns represent weights feeding into specific units. Rows represent the output connectivity of individual units. Larger weights are represented by larger boxes. Positive weights are depicted by white boxes, negative weights by black ones. This network was trained on 'ABA' sentences. Regions labeled 1 show weights that encode relationships between features of A syllables. Relationships between features of B syllables are identified by regions identified as 2. Correlations between features of A and B words are identified in regions labeled 3. Weights that connect the duplicated features of both A syllables are shown in regions numbered 4.*

That is, units 1–4 encoded the first A word and units 9–12 the second, identical A word.

The weight matrix obtained through learning is essentially an approximation of the correlation matrix between all features of the three 'syllables' over all 16 'sentences'. What the network learns is thus not independent of the specific features of the tokens that were used. However, these are just arbitrary and random. At the end of habituation to 'ABA' sentences for example, what the network has learned can be broken down into four distinct components (identified by corresponding numbered regions in Figure 5).

*First*, to what extent are the activations on different features of A 'syllables' correlated with one another (all pairwise correlations within units 1–4, within units 9–12, and between units of banks 1–4 and 9–12 except for corresponding units 1 and 9, 2 and 10, 3 and 11, and 4 and 12)? These correlations capture the similarity of the four A 'syllables' in the training set. For example, this network learned that the first and second features of A 'syllables' were negatively correlated.

*Second*, to what extent are activations on features of B 'syllables' are correlated with one another (all pairwise correlations within units 5–8)? This refers to the similarity of the four B 'syllables' the network was presented. This network learned, for example, that the first and second features of B 'syllables' were also negatively correlated.

*Third*, to what extent are activations on features of A 'syllables' correlated with B 'syllables' (all pairwise correlations between units 1–4 and 5–8, and between units 5–8 and 9–12)? This captures the similarity between A and B 'syllables' in the training set. For this network, the first features of A and B 'syllables' were negatively correlated, as were the third and fourth features. This is reflected in the network's weights.

*Fourth*, the network has knowledge that activations on corresponding units for duplicate A 'syllables' are perfectly correlated (correlations for corresponding units 1 and 9, 2 and 10, 3 and 11, and 4 and 12). This last piece of knowledge is crucial because novel 'syllables' in the test phase will be dissimilar (to various random degrees) with all three other pieces of knowledge, but not with the latter one for consistent test sentences. A and A′ syllables may be considerably different (where A and A′ are phonemes used in A slots in habituation and testing phases, respectively), and so may B and B′ syllables. The similarity between A and B could be different than that between A′ and B′. Whereas the network learns something specific from the restricted training set, there is one source of knowledge that allows it to distinguish between consistent and inconsistent 'sentences' even when novel items are used. This was made possible by learning that corresponding features of both A 'syllables' are identical.[8]

From a psychological perspective, at the end of the habituation phase, networks have learned what the prototypical A syllable is, what the prototypical B syllable is, to what extent A and B syllables are related, and the temporal organization of these prototypical A and B syllables. Most importantly, the network will also have learned that corresponding features of the duplicated A syllable are identical. When testing begins, A′ and B′ syllables will differ from the prototypical A and B, respectively. The similarity between A′ and B′ syllables may differ from the similarity between A and B syllables. This will affect consistent and inconsistent test sentences equally, on average. What the network learned about the corresponding features in redundant syllables is critical and is not affected by the introduction of novel syllables. This is what allows networks to settle more quickly on consistent than on inconsistent test sentences.

## Discussion

The networks in this simulation of Marcus *et al.*'s (1999) Experiment 1 capture what these authors observed in 7-month-olds. Networks habituated to the initial structure by learning stable representations. When tested with novel patterns organized in consistent and inconsistent structures, they required significantly more pattern presentations to learn inconsistent items. Marcus and colleagues (1999) reported that a few infants actually attended longer to consistent items than to inconsistent ones; this was observed in a few of our networks as well. Despite Marcus's (1999) claim to the contrary, networks without rules or variables trained on distributed binary input can capture the underlying structure in their learning environment and generalize to novel items. In fact, the weight diagram in Figure 5 represents a computationally precise form of prototype. Weights represent both the underlying structure in all sentences and the specifics of the random set of habituation items selected from Table 1.

It is worth noting that our coding scheme, distinct from a coding scheme based on continuous features represented on fewer units (e.g. Negishi, 1999; Shultz,

[8] To support this conclusion, we examined the performance of networks in which either we reset these corresponding weights to zero after habituation and before testing or we prevented weights for corresponding features from changing during habituation and testing phases. In both cases, networks exhibited no significant difference in number of test cycles for consistent and inconsistent test patterns, although a significant effect of block was observed (all $p$ values were greater than 0.05, except for block where $p$ values were less than 0.05).

1999), creates spurious relationships in the habituation phase. Randomly selecting four patterns as 'A syllables' and four patterns as 'B syllables' from Table 1 for the habituation phase results in spurious within-syllable correlations. By using a distributed encoding scheme, we assume that all syllables share various levels of similarity with one another. Given any four patterns, the correlation between the activity of any two units representing these patterns ranges between −1 and 1. Such correlations will be reflected in the weights of the network at the end of habituation training. In Figure 5, these weights can be seen in regions 1, 2 and 3. The spurious correlations within the test patterns (also randomly selected from Table 1) will, on average, be inconsistent with what the network has learned during habituation, and will therefore slow down processing. However, such an effect will occur for both consistent and inconsistent test patterns, and therefore cannot be the source of any reliable differences observed between the processing of consistent and inconsistent test patterns.

The fact that networks required significantly more processing on inconsistent test items emphasizes the importance of the underlying statistical structure across the habituation sentence as a whole rather than within individual syllables. The corresponding weights (region 4 in Figure 5) are the largest between both A syllables, and are not affected by within-syllable correlations. When testing begins with novel syllables, networks will anticipate that the third syllable is the same as the first, although activations on the units encoding the last syllable will be somewhat distorted from the within-syllable correlations learned in the habituation phase (i.e. the two regions identified as 1 in the bottom-right corner of Figure 5).

We stress that the performance of our model is based entirely on statistical learning. There are no rules applied to variables involved in the performance of the networks. Marcus (1999) has argued that neural networks that use a continuous encoding scheme on local units actually implement variables, and therefore weights implicitly implement rules. Although we do not agree with this argument,[9] it cannot be applied to our distributed, binary encoding scheme.

[9] For variable bindings to be useful, they have to be preserved and accessible to further computation, as is the case in explicit variable-binding schemes. In networks with hidden units between input and output units, assignments of analog values to inputs are lost as soon as activation is propagated onto nonlinear hidden units (Shultz, 1999). But more generally, the notion of variables is only meaningful in the context of a system that explicitly manipulates variables. Even if the input to a network is analog, and the network behaves in a rule-like way, this does not mean that the network performs symbolic computations on rules and variables. Indeed, the networks in question do not use rules and variables.

### *Simulation 2: AAB versus ABB*

Before moving on to the general discussion, we report on a second simulation aimed to model Marcus *et al.*'s (1999) Experiment 3. In this experiment, which used 'AAB' and 'ABB' structures, habituation patterns and both types of test patterns involved a consecutive duplication of one element (i.e. 'AA' or 'BB'). That is, both 'grammar' types involve immediate temporal duplication, whereas this was not the case for the 'ABA' grammar in Marcus *et al.*'s (1999) Experiments 1 and 2.

#### Method

Thirty networks were trained and tested in this simulation. The procedure was identical to Simulation 1, except that the training patterns were constructed so as to follow an 'AAB' structure. Consistent test items involved novel syllables organized into 'AAB' patterns, whereas inconsistent test items followed an 'ABB' structure.

#### Results

The networks required an average of 23.4 epochs to learn in the habituation phase (SD = 2.81). No network

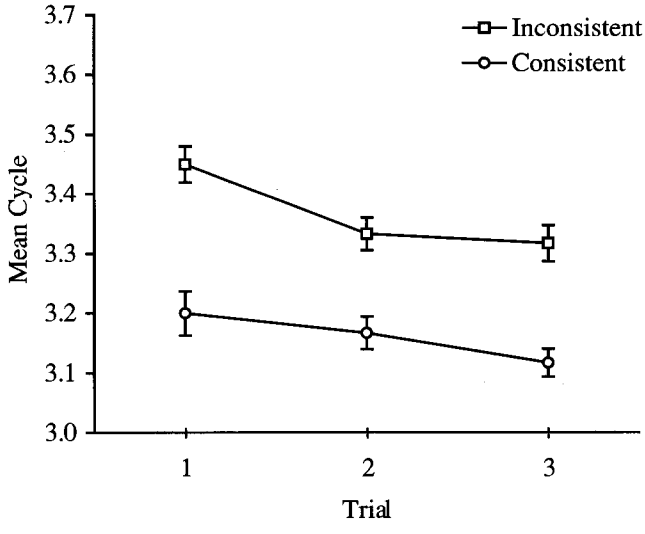


**Figure 6** *Mean number of testing cycles required for network learning of novel consistent (AAB) and inconsistent (ABB) test patterns by trial in Simulation 2. Error bars show standard deviations.*

reached the 50 epochs limit. Figure 6 depicts the number of testing cycles to learn consistent and inconsistent test items over testing blocks. The average number of pattern presentations for consistent items was 3.2, 3.1 and 3.1 for test blocks 1, 2 and 3. For inconsistent items, the average number of testing cycles was 3.4, 3.3 and 3.3, respectively. Four of the networks required more pattern presentations for consistent patterns, on average, than for inconsistent patterns. Testing cycles required to learn test items were analyzed with a 2 by 3 by 2 repeated-measures analysis of variance, with type (consistent and inconsistent), block (1–3) and pattern (first and second) as within-subject factors. The analysis yielded a significant effect of type ($F(1, 29) = 13.05$, $p < 0.05$) as well as a significant effect of block ($F(2, 58) = 5.23$, $p < 0.05$). There was no significant effect of pattern, nor any significant interaction. The tests of within-subject contrasts showed a significant linear trend for block ($F(1, 29) = 6.99$, $p < 0.05$).

Figure 7 shows the weight diagram of a network that required 23 habituation epochs and required less processing for consistent test items. These are weights recorded at the end of the habituation phase. This diagram should be interpreted as Figure 5 was, with the exception that units 5–8 encode the second A syllable and units 9–12 encode the B syllable. Networks at the end of the habituation phase anticipate the second A syllable to be identical to the first A syllable. This can be seen in the square section delimited by rows 1–4 and columns 5–8. Before the second syllable is introduced, units encoding this syllable will become active in a way that reflects activity on corresponding units from the first syllable.

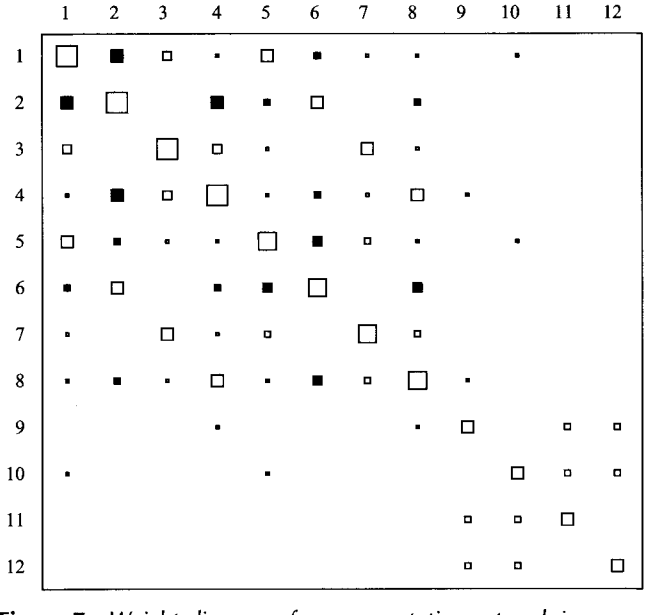


**Figure 7** *Weight diagram of a representative network in Simulation 2. This network was trained on 'AAB' sentences.*

### Discussion

Networks in this second simulation capture the results of Marcus *et al.*'s (1999) Experiment 3. Auto-associator networks habituated to the initial patterns and showed differential recovery to consistent and inconsistent test items. In this simulation, as in the previous one, networks generalized what they learned about the underlying structure in the habituation set to patterns with novel syllables. In both cases, networks capture what Marcus *et al.* (1999) and Pinker (1999) argued can only be captured by rules with variables.

Another way of characterizing our networks' performance is to say that they learned a type based on a restricted set of tokens, and that they can generalize their knowledge of type to novel tokens. This is one way in which symbol-processing systems process information. However, there are no rules, symbols, types or tokens in our networks. This shows that some of the functionality of symbol-processing systems naturally emerges from simple, statistical mechanisms.

## General discussion

Our simulations captured the empirical data of Experiments 1 and 3 reported in Marcus *et al.* (1999). Networks habituated to the training patterns by forming stable representations of the input. When presented with test patterns consisting of novel elements, they required significantly more processing on patterns that were inconsistent with the structure underlying the habituation patterns. We did not report simulations of Marcus *et al.*'s (1999) Experiment 2, which controlled for patterns of phonetic features in their Experiment 1, because their results showed that the pattern of phonetic features did not matter.

We suggested that the networks implemented a computational equivalent of the prototype interpretation of habituation data (Younger & Cohen, 1985). A prototype is formed by averaging perceptual features of the various stimuli presented (Rosch, 1978). Infants can extract more than one prototype when the familiarization stimuli represent more than one class of items (Younger & Cohen, 1985). The important point, however, is that prototypes are constructed by implicit computations based directly on the perceptual features of the stimuli. This is different from transforming perceptual input into variable values upon which

symbolic computations may be carried out. This symbolic position is advocated by Marcus *et al.* (1999), and it requires additional steps compared to averaging perceptual input (namely, redescribing the input in symbolic terms and then performing rule-based computations). Our position, which is that of Sokolov (1963) as it turns out, is that a model based directly on the perceptual input can account for the data, without requiring this input to be translated in symbolic form for algebraic computations.

Does this mean that the prototype interpretation of habituation, implemented by the networks, is psychologically correct? As long as equally successful alternative accounts can be entertained, the answer is no. What the success of our simulations suggests, however, is that the conclusion that 7-month-old infants learn and use rules as explicit computational mechanisms (Marcus, 1999; Marcus *et al.*, 1999; Pinker, 1999) is at best premature. Contrary to previous claims, statistical learning devices *can* capture these empirical regularities.

Our model requires that the same banks of units are used for individual syllables, and the distributed coding scheme further implies similarity between items. The latter is an important assumption. McClelland and Plaut (1999) discussed several 'natural' encoding schemes of the phonemes used by Marcus and his colleagues (1999) that also imply similarity between phonemes.

Similarity and identity are not incompatible, as similar items (e.g. two dogs) can nevertheless have a unique identity. Marcus *et al.*'s (1999) interpretation can ignore similarity because it is based on identity (i.e. rules such as $y = f(x)$ would not be affected if the various $x$ values were similar). Unless it is shown that phonemes share no similarity, and McClelland and Plaut (1999) have argued against that, our encoding scheme is not what distinguishes our model from Marcus *et al.*'s (1999) interpretation. Abstract algebraic rules could be successfully applied to our patterns. The point, however, is that statistical procedures can also capture the infants' performance.

Marcus and colleagues (1999) acknowledged experiments that showed a capacity in infants to learn transitional probabilities (e.g. Saffran *et al.*, 1996). Because they assumed that statistical regularities could not be learned in their own experiment, Marcus *et al.* (1999) argued that infants possess at least two learning mechanisms – a statistical learning mechanism sensitive to transitional probabilities and a mechanism of algebraic rules for problems such as in their experiments. What is missing in this dual-process account is an additional mechanism that knows which learning mode should be selected for any particular task.

A possible solution to this problem is found in Marcus (1999), where he suggests that distributed binary encoding is appropriate for capturing transitional probabilities. We have demonstrated that this coding scheme allows networks to capture underlying structures such as 'ABA' or 'ABB' and generalize to novel items; it may thus be that our auto-associator model provides an integrated account of both types of habituation data. It is both simpler and more consistent with other findings than the account provided by Marcus and colleagues (1999). Until unequivocal evidence of rule learning in 7-month-old infants is reported, the interpretation found in Marcus *et al.* (1999) and Pinker (1999) is not as parsimonious.

This is related to the broader issue of interpreting habituation data (Fischer & Bidell, 1991, 1992). Simpler perceptual-level accounts have been offered that question suggestions of conceptual knowledge, reasoning, rules, surprise and so on in infants (e.g. Bogartz *et al.*, 1997; Cohen, 1998; Haith, 1998; Sirois *et al.*, in preparation). The data obtained in these experiments are as valuable as any other data, and our paper is one of the many that testify to the heuristic value of habituation experiments that investigate infant competence. However, until a definitive demonstration of rules, concepts or conceptualized expectation is made in young infants, we would argue in favor of parsimony. Perceptual-level accounts should be preferred because they are simpler and warranted by the procedure.

In a recent paper, Gomez and Gerken (1999) report findings from experiments similar to those of Marcus and colleagues (1999). Also using a familiarization procedure (in which habituation was assumed rather than measured and individual patterns were not presented repeatedly in the habituation phase), 1-year-old infants were trained on speech streams generated by a finite-state grammar. In four experiments, infants showed a preference for consistent test items over inconsistent items. That is, infants attended longer to novel, consistent items than to novel, inconsistent ones. This could be problematic for Marcus and colleagues (1999), and consequently for our model, where novelty preference rather than familiarity preference is observed.

The stimuli used by Gomez and Gerken (1999) are more complex than the ones used by Marcus *et al.* (1999). In both cases, the data suggest that infants distinguish between consistent and inconsistent items. However, the familiarity preference observed by Gomez and Gerken (1999) could reflect incomplete learning at the end of the habituation phase (Hunter, Ross & Ames, 1982). Bogartz and colleagues (1997) suggest that infants with partial knowledge could (implicitly) distinguish consistent and inconsistent stimuli, yet ignore the latter in favor of appropriately learning the former. When habituation is assumed rather than measured, the extent of learning

cannot be assessed. Although infants obviously distinguished both types of items, insufficient learning could have prevented them from exhibiting the novelty preference observed with Marcus and colleagues' (1999) simpler stimuli. Both sets of data may well be compatible. Whether our auto-associator model could capture Gomez and Gerken's (1999) data is beyond the scope of this paper. Coverage of their data is conceivable because there are perceptual regularities in their stimuli such as those that were crucial in the present simulations.

A final issue worth noting is how our auto-associator model compares with other neural network models of habituation. Kohonen (1988) proposed the *novelty filter* as a neural network model of habituation. Although both our model and Kohonen's use the auto-associator architecture, the learning rules are quite different. At the end of habituation training, the novelty filter no longer responds to the habituation items. That is, if a pattern was part of the habituation set, no unit in the network will be active when it is presented in the test phase. Units in the network will be active only when patterns with novel features are presented. It follows that such a model would capture the novelty effect (i.e. recovery) in typical habituation experiments.

The differences between Kohonen's (1988) model and the one we present in this paper lead to an empirical prediction. Our networks can distinguish between the different habituation items at the end of training. That is, the output of these networks can serve as useful, discriminable input to other networks. The novelty filter on the other hand returns an identical output for all familiar items; its only use is, as the name implies, to detect novelty. If infants were empirically shown to distinguish between habituated items, the novelty filter would not serve as a good model.

Our auto-associator approach is different from simple recurrent network (SRN) models, which are explicitly trained to predict the value of the next item in sequential tasks. SRNs usually employ hidden units, allowing them to learn complex nonlinear relationships. Predictive ability in our networks would be the outcome of learning simple, first-order statistics between sequential input values, and not a consequence of training geared towards prediction. Moreover, our networks can learn relationships between sequentially distant units. This allows our auto-associator networks to learn relationships that would elude an SRN. For example, the four vectors $\{(-1\ -1\ 1)(1\ -1\ -1)(-1\ 1\ 1)(1\ 1\ -1)\}$ could be reproduced in a sequential auto-associator, which would learn that items 1 and 3 are negatively correlated, but an SRN could not learn to predict the third item of individual vectors because the second item is unrelated to the third item.

Another important difference between the auto-associator model and recurrent networks is in how we have represented time. The temporal window in SRNs is usually limited to discrete items, because the task is one of predicting the next item. Such models of Marcus *et al.*'s (1999) experiments are essentially performing word parsing. Our model, on the other hand, performs a form of sequential sentence parsing. That is, the temporal window is larger. SRNs learn to predict the next word, whereas our networks learn to represent the whole sentence, albeit sequentially. Which approach is more appropriate is an empirical question.

Mareschal and French (1997, 2000) presented the auto-encoder feedforward architecture as a model of habituation. As in our auto-associator model, such networks learn to reproduce the input they receive. However, the input is reproduced on a bank of output units, and a set of hidden units mediates the propagation of activation from input to output units. The number of hidden units is less than the number of units required to encode patterns on input and output units, and thus networks must abstract information from the input in a more compact form in order to reproduce it on output units (Mareschal & Thomas, in press).

The auto-encoder model shares important features with a variety of neural network models of Marcus *et al.*'s (1999) data. Several feedforward architecture models reportedly capture the distinction between grammatically consistent and inconsistent items (Christiansen & Curtin, 1999; Negishi, 1999; Seidenberg & Elman, 1999; Shultz, 1999). The details of each model, and their relative success at modeling the Marcus *et al.* (1999) data, would deserve a paper in themselves. For the purposes of the current paper, we highlight two key distinctions between these models and our auto-associator networks.

First, these feedforward models of Marcus *et al.*'s (1999) data all make an assumption about capturing the temporal nature of recovery data. Network error on the various test patterns is used as an index of a model's ability to distinguish between classes of test items. However, it is not clear that this error index will translate into an analog of infant looking. Sirois and Shultz (1998) discussed how larger network error does not necessarily lead to additional processing; in some circumstances, larger error can actually make learning faster. In the auto-associator simulations we presented in this paper, the processing index for test items is temporal in nature, providing a more direct mapping of the empirical procedure.

Second, most of these feedforward models of Marcus *et al.*'s (1999) data make use of hidden units between input and output. Such networks can learn complex

nonlinear functions. However, networks with a similar level of architectural complexity are also used to model cognitive processes in older children (e.g. Elman *et al.*, 1996). In the context of the debate over interpretations of habituation data (e.g. Bogartz *et al.*, 1997; Haith, 1998), one question that follows is whether such powerful networks are necessary to model habituation in young infants. Our model highlights that the ability to represent simple linear statistics is sufficient.

To summarize, our simulations of Marcus *et al.*'s (1999) habituation experiments successfully capture the reported empirical regularities. This supports an alternative, perceptual perspective to these authors' rule-based interpretation. We suggest that there is no unequivocal support for abstract, algebraic rules in 7-month-olds. Future work aimed at contrasting the various neural network models of habituation may prove fruitful for the general goal of explaining infant competence.

## Acknowledgements

This research was supported in part by a Natural Sciences and Engineering Research Council of Canada (NSERC) graduate fellowship to the first author, and an NSERC operating grant to the third author. The authors thank Alan Bale, Louise Charrois, Jacques Katz, Yuriko Oshima-Takane and François Rivest for their helpful suggestions. The authors also wish to thank the anonymous reviewers of earlier versions of the manuscript for their useful comments.

## References

Anderson, J.A. (1977). Neural models and cognitive implications. In D. LaBerge & S.J. Samuels (Eds), *Basic processes in reading: Perception and comprehension* (pp. 27–90). Hillsdale, NJ: Erlbaum.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review*, **84**, 413–451.

Baillargeon, R. (1987). Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology*, **23**, 655–664.

Baillargeon, R., Spelke, E.S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, **20**, 191–208.

Bogartz, R.S., Shinskey, J.L., & Speaker, C.J. (1997). Interpreting infant looking: the event set × event set design. *Developmental Psychology*, **33**, 408–422.

Christiansen, M.H., & Curtin, S.L. (1999). The power of statistical learning: no need for algebraic rules. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 114–119). Hillsdale, NJ: Erlbaum.

Clifton, R.K., & Nelson, M.N. (1976). Developmental study of habituation in infants: the importance of paradigm, response system, and state. In T.J. Tighe & R.N. Leaton (Eds), *Habituation: Perspectives from child development, animal behavior and neurophysiology* (pp. 159–205). Hillsdale, NJ: Erlbaum.

Cohen, L.B. (1998). An information-processing approach to infant perception and cognition. In F. Simion & G. Butterworth (Eds), *The development of sensory, motor and cognitive capacities in early infancy* (pp. 277–300). Hove, UK: Psychology Press.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fischer, K.W., & Bidell, T.R. (1991). Constraining nativist inferences about cognitive capacities. In S. Carey & R. Gelman (Eds), *The epigenesis of mind: Essays on biology and cognition* (pp. 199–235). Hillsdale, NJ: Erlbaum.

Fischer, K.W., & Bidell, T.R. (1992). Ever younger ages: constructive use of nativist findings about early development. *SRCD Newsletter*, Winter Issue, 1–14.

Gomez, R.L., & Gerken, L.A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, **70**, 109–135.

Haith, M.M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, **21**, 167–179.

Hunter, M.A., Ross, H.S., & Ames, E.W. (1982). Preferences of familiar or novel toys: effect of familiarization time in 1-year-olds. *Developmental Psychology*, **18**, 519–529.

Kemler, D.G. (1981). New issues in the study of infant categorization: a reply to Husaim and Cohen. *Merrill-Palmer Quarterly*, **27**, 457–463.

Kohonen, T. (1977). *Associative memory: A system theoretical approach*. New York: Springer.

Kohonen, T. (1988). *Self-organization and associate memory*. New York: Springer.

Malcuit, G., Pomerleau, A., & Lamarre, G. (1988). Habituation, visual fixation and cognitive activity in infants: a critical analysis and attempt at a new formulation. *European Bulletin of Cognitive Psychology*, **8**, 415–440.

Marcus, G.F. (1999). Do infants learn grammar with algebra or statistics? *Science*, **284**, 433.

Marcus, G.F., Vijayan, S., Bandi Rao, S., & 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 M.G. Shafto & P. Langley (Eds), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 484–489). Hillsdale, NJ: Erlbaum.

Mareschal, D., & French, R.M. (2000). Mechanisms of categorization in infancy. *Infancy*, **1**, 51–67.

Mareschal, D., & Thomas, M.S.C. (in press). Self-organization in normal and abnormal cognitive development. In A.F. Kalverboer & A. Gramsbergen (Eds), *Brain and behaviour in*

*human development: A source book*. Amsterdam: Kluwer Academic.

Mareschal, D., Plunkett, K., & Harris, P. (1995). Developing object permanence: a connectionist model. *Proceedings of the 17th Annual Meeting of the Cognitive Science Society* (pp. 170–175). Hillsdale, NJ: Erlbaum.

McClelland, J.L., & Plaut, D.C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Science*, **3**, 166–168.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**, 159–188.

Mix, K.S., Levine, S.C., & Huttenlocher, J. (1997). Numerical abstraction in infants: another look. *Developmental Psychology*, **33**, 423–428.

Munakata, Y., McClelland, J.L., Johnson, M.H., & Siegler, R.S. (1997). Rethinking infant knowledge: toward an adaptive process account of success and failures in object permanence tasks. *Psychological Review*, **104**, 686–713.

Negishi, M. (1999). Do infants learn grammar with algebra or statistics? *Science*, **284**, 433.

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

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Seidenberg, M.S., & Elman, J.L. (1999). Do infants learn grammar with algebra or statistics? *Science*, **284**, 433.

Shultz, T.R. (1999). Rule learning by habituation can be simulated in neural networks. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 665–670). Hillsdale, NJ: Erlbaum.

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

Sirois, S., Debbané, M., & Zelazo, P.R. (in preparation). Habituation and cognition: growing concerns. Manuscript in preparation.

Sokolov, E.N. (1963). *Perception and the conditioned reflex*. New York: Macmillan.

Spelke, E.S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, **21**, 181–200.

Spelke, E.S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, **99**, 605–632.

Starkey, P., Spelke, E.S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, **36**, 97–127.

Thorpe, W.H. (1963). *Learning and instinct in animals*, 2nd edn. Cambridge, MA: Harvard University Press.

Ungerleider, L.F. (1995). Functional brain imaging studies of cortical mechanisms for memory. *Science*, **270**, 769–775.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, **358**, 749–750.

Wynn, K. (1995). Infants possess a system of numerical knowledge. *Current Directions in Psychological Science*, **4**, 172–177.

Younger, B.A., & Cohen, L.B. (1985). How infants form categories. *The Psychology of Learning and Motivation*, **19**, 211–247.

Zelazo, P.R. (1988). Infant habituation, cognitive activity and the development of mental representations. *European Bulletin of Cognitive Psychology*, **8**, 649–654.

Received: 9 June 1999
Accepted: 27 March 2000