A Critical Look at the Mechanisms Underlying Implicit Sequence Learning

Todd M. Gureckis (gureckis@psy.utexas.edu)
Bradley C. Love (love@psy.utexas.edu)
Department of Psychology; The University of Texas at Austin
Austin, TX 78712 USA

Abstract

In this report, a model of human sequence learning is developed called the linear associative shift register (LASR). LASR uses a simple error-driven associative learning rule to incrementally acquire information about the structure of event sequences. In contrast to recent modeling approaches, LASR describes learning as a simple and limited process. We argue that this simplicity is a virtue in that the complexity of the model is better matched to the demonstrated complexity of human processing. The model is applied in a variety of situations including implicit learning via the serial reaction time (SRT) task and statistical word learning. The results of these simulations highlight commonalities between different tasks and learning modalities which suggest similar underlying learning mechanisms.

Introduction

One of the most striking aspects of human behavior is the ease with which we can acquire new skills with little conscious effort. In order to better understand this phenomena, a large literature has developed exploring the ability of participants to implicitly learn about the sequential structure of a series of events (see Cleeremans, Destrebecqz, Boyer, 1998, for a review). However, the type of memory and learning mechanisms which might support such learning are not well understood (see Keele, Ivry, Mayr, Hazeltine, & Heuer, 2004 or Sun, Sluzarz, & Terry, 2005 for some recent proposals).

In this paper, we develop a simple model of sequence learning behavior called the linear associative shift-register (LASR). The model is unique from past approaches in that it describes implicit sequence learning as a simple and limited process which operates on a small temporary buffer of past events. This contrasts with other models of sequence learning which have described learning as a more complex and flexible process (Cleeremans & McClelland, 1991; Cleeremans, 1993; Lebiere & Wallach, 2000).

There are two main goals of this report. First, we demonstrate how a very simple learning mechanism such as LASR can provide a detailed account of a number of findings from the implicit sequence learning literature. A key criticism we develop is that in previous modeling accounts (such as the simple recurrent network (SRN) of Cleeremans, 1993), the complexity of the model is not well matched to the demonstrated complexity of the learner. While LASR cannot explain all aspects of our rich sequential behavior, we believe the model provides a unique baseline against which to test more complex theories and experiments.

Second, we explore the ability of this simple model to account for sequential learning phenomena in a variety of implicit learning situations including the serial reaction time (SRT) task and statistical word learning paradigms. LASR provides a similar account of the type of processing which underlies performance in both kinds of tasks, suggesting that they may rely on similar underlying mechanisms.

We begin by introducing the LASR model and the principles upon which it is based. Next, we consider a study conducted by Lee (1997) assessing implicit learning of sequentially structured material. Finally, we explore the ability of LASR to account for statistical word learning in infants as reported by Saffran, Aslin, and Newport (1996).

The Linear Associative Shift-Register (LASR) Model

LASR is a mechanistic model of implicit sequence learning. The model describes implicit sequence learning as the task of appreciating the associative relationship between past events and future ones. LASR assumes that subjects maintain a limited memory for the sequential order of past events and that they use a simple error-driven associative learning rule (Widrow & Hoff, 1960; Rescorla & Wagner, 1972) to incrementally acquire information about sequential structure. Despite its simplicity, the model can very quickly learn to appreciate rather complex dependencies between events which are structured in time. The model is organized around 3 principles:

1. Past events are stored in a temporary buffer The model begins by assuming a simple shift-register memory for past events. Individual elements of the register are referred to as *slots*. New events encountered in time are inserted at one end of the register and all past events are accordingly shifted one time slot. Thus, the most recent event is always located in the right-most slot of the register (see Figure 1). This form of memory maintains the sequential order of recent events using spatial position (similar to many other models, see Sejnowski and Rosenberg (1987) or Cleereman's (1993) buffer network).

2. Learning to predict what comes next This simple memory mechanism forms the basis of a *detector* (see Figure 1). A detector is a simple, single-layer linear network or perceptron (Rosenblatt, 1958) which learns to predict the occurrence of a single future event based

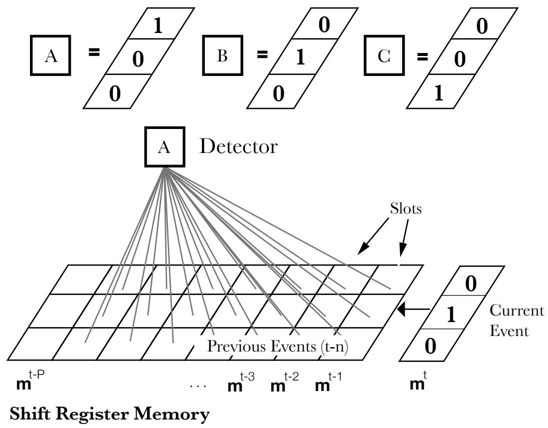

Figure 1: A shift-register memory and a single detector. New events encountered enter into the register from the right and are stored in the sequence they arrived in the memory register.

on past events. Because each detector predicts only a single event, a separate detector is needed for each possible event. Each detector has a weight from each event outcome at each time slot. On each trial, activation from each memory-register slot is passed over a connection weight and summed to compute the activation of the detector's prediction unit. The task of a detector is to adjust the weights from individual memory slots so that it can successfully predict the future occurrence of it's response. Each detector learns to strengthen the connection weights for memory slots which prove predictive of the detector's response while weakening those which are not predictive or are counter-predictive.

3. Recent events have more influence on learning than past events The model assumes that events in the recent past are remembered better than events which happened long ago. This effect is implemented by attenuating the activation strength of each register position by how far back in time the event occurred. Because of this, an event which happened at time $t-1$ has more influence on future predictions than events which happened at $t-2$, $t-3$, etc... Similarly, learning is slower for slots which are positioned further in the past because their activation strength is reduced (see Equation 4).

Model Formalism

The following section describes the mathematical formalism of the model. The model is easily described using three equations and three intuitive parameters.

Memory As illustrated at the top of Figure 1, input to the model on each time step is a N-dimensional vector $\mathbf{m}^t$ where each entry m_i^t corresponds to the presence ($m_i^t = 1$) or absence ($m_i^t = 0$) of event i on the current trial, t. The complete history of past events is thus a NxP matrix, $\mathbf{M}$, where N is the number of possible events, and P is the number of events so far experienced and stored in memory. The shift-register memory of past events is indexed based on the current time t. Thus, $\mathbf{m}^{t-1}$ refers to the input vector experienced on the previous time step, and $\mathbf{m}^{t-2}$ refers to the input experienced two time steps in the past.

Response Given N possible events or choice options, the model has N detectors. The activation d_k of the detector k at the current time, t, is computed as the weighted sum of all events in all slots multiplied by an exponential attenuation factor:

$$d_k = \sum_{i=1}^{P} \sum_{j=1}^{N} w_{(t-i)jk} \cdot m_j^{t-i} \cdot e^{-\alpha \cdot (i-1)} \tag{1}$$

where $w_{(t-i)jk}$ is the weight from the jth outcome at time slot $t-i$ to the kth detector, and $m_j^{t-i} \cdot e^{-\alpha.(i-1)}$ is the outcome of the jth option at time $t-i$ multiplied by the memory attenuation factor. The α is a free parameter which controls the rate of decay for traces in memory. The final output of each detector, o_k, is a sigmoid transform of the activation, d_k, of each detector:

$$o_k = \frac{1}{1+e^{d_k}}. \tag{2}$$

When being compared to human data, the output of each detector is converted into a response probability or tendency (p_k) using the Luce choice rule (Luce, 1959):

$$p_k = \frac{o_k}{\sum_{j=1}^{N} o_j} \tag{3}$$

For example, human reaction time is assumed to inversely relate to this response tendency so that faster responses in the task correspond to higher values of p_k (Cleeremans & McClelland, 1991).

Learning Learning in the model is implemented using the well known delta-rule for training single layer networks (Widrow & Hoff, 1960) with a small modification introduced by Rumelhart and McClelland (1986) (sometimes referred to as the generalized delta-rule for single layer networks). For each detector, the difference between the actual outcome of the current trial, t_k, and the output of the detector, o_k, is computed and used to adjust the weights:

$$\Delta w_{ijk} = \eta \cdot (t_k - o_k) \cdot m_j^i \cdot e^{-\alpha \cdot (i-1)} \cdot d_k(1-d_k) \tag{4}$$

The Δw_{ijk} value is added to the corresponding weight after each learning episode. The η is a learning rate parameter and $e^{-\alpha \cdot (i-1)}$ is the memory attenuation factor described above and $d_k(1-d_k)$ is a factor representing the derivative of the sigmoid transfer function with respect to the weights which moves learning on each trial in the direction of gradient descent on the error. In the simulations reported here $\alpha = 0.2$ and $\eta = 0.9$.

Evaluating the LASR model

In the following section we explore the ability of this model to account for a number of published findings concerning implicit sequence learning. The results illustrate

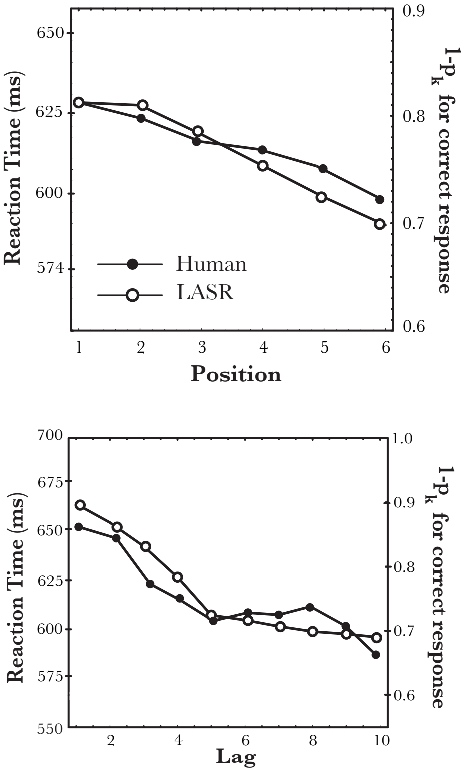

Figure 2: *Top:* Human reaction time and model response as a function of set position in Boyer, Destrebecqz, and Cleeremans' (1998). *Bottom:* Human and model response as a function of the lag separating two occurrences of the same event. All human data replicated approximately from figures in Boyer, et al. (1998)

how the simple principles which define the LASR model are able to provide a strong account of learning and show the relationship between the data collected across a number of paradigms.

Sequence Learning via the SRT task

The majority of SRT studies have used simple repeating sequences of various lengths. One notable exception is Lee (1997). In this study, the pattern of stimuli was determined by a simple, yet subtle rule: each of six choice options had to be visited once in each set of six trials in a random order. Examples of legal six-element sequence sets are 132546, 432615, and 546123. Boyer, Destrebecqz, and Cleeremans (1998) provide a replication of Lee (1997) and showed that reaction time monotonically decreases as a function of set position 1-6 (see Figure 2, top panel).

What is unique about the sequence employed by Lee (1997) is that while it is generated by a simple rule, each stimulus item can be followed by any other item. The key predictive structure is contained in the set of six successive elements which avoid repetition. Can the simple one-layer associative learning mechanism in LASR account for such a result?

Simulation Results LASR was applied to the task in a similar manner to how participants were trained with the same number of trials and the same sequential structure as the Boyer, Destrebecqz, and Cleeremans (1994) replication. On each trial, the magnitude of the model's response for the correct outcome was recorded. Figure 2 (top panel) shows the model's response as a function of position. At the first set position, the model's error is about 0.83 which is chance (i.e, 5/6) but as more of the sequence is revealed, the model continues to reduce this error (thus predicting faster RT).

The model is able to replicate the key qualitative results of the study despite having no mechanism for grouping sequence elements, and only a simple single layer of weights. A closer look at how the model solves the problem gives some insight into the structure of the task. Figure 3 shows the setting of each of the weights in the model at the end of learning. The key pattern to notice is that the diagonal entries for each past time slot are strongly negative while all other weights are close to zero. The diagonal of each weight matrix represents the weight from each event to it's own detector. Thus, the model attempts to inhibit any response that occurred in the last few trials.

The impact of this is demonstrated in Figure 2 (bottom panel) which shows response probability as a function of the number of events separating two repeated events (lag). Since the same event could not repeat on successive trials, repeated events were at minimum separated by 1 event (lag-1). This might happen if the fifth event of one sequence repeated as the first element of the next sequence. Figure 2 (bottom) shows that as the lag between two repeated events increases, the model accurately predicts faster RT. The memory attenuation of past events causes them to become less inhibited as they move further into the past (i.e., events at lag-10 are more strongly inhibited than events at lag-1). Boyer, et al. (1998) examined this same lag effect in the reaction time of participants in their replication and found an identical effect (also shown in Figure 2, bottom panel). Participant RT was inversely related to the number of trials that separated the repeated event. The model describes performance in the task as a simple negative recency effect.

Boyer, et al. (1998) explored how the SRN accounted for human performance in this task. The SRN provides a similar conceptual account by learning to increase the likelihood of an response as a function of the number of events since last experienced. However, the learning mechanism of the SRN is much more complicated than that of LASR because the model must acquire appropriate hidden unit representations in addition to adjusting weights in the upper layer of the network. As a result, Boyer, et al. (1998) had to train the SRN on considerably more trials than humans or LASR. Both humans and LASR were trained for 4320 trials (24 blocks of 180 trials each), whereas the SRN was trained for 30,240 trials. In addition, the hidden unit representations the SRN acquires are difficult to interpret because the model learns to predict successors of particular aggregate con-

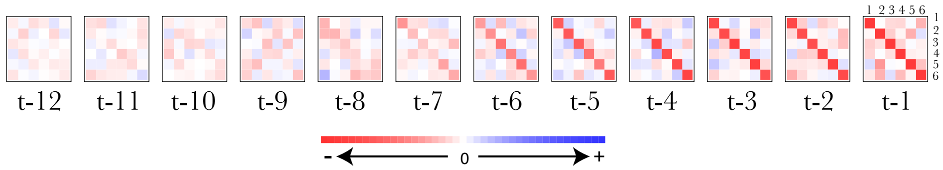

Figure 3: The final LASR weights for the Lee (1997) sequence learning problem. Negative weights are darker red. Positive weights are darker blue. The weights leaving each memory slot ($t-1$, $t-2$, etc...) are shown as a separate matrix. Each matrix shows the weights from each stimulus element to each detector. For example, the red matrix entry in the top left corner of t-1 slot is the weight from event "1" to the detector for event "1".

texts. Instead, LASR clearly describes performance in the task as a simple negative recency effect, where recent events are inhibited.

Boyer, et al. (1998) point out that in both their replication and in the original Lee (1997) study, participants demonstrate a faster reaction time to the latter elements of the each sequence even in the first block of learning and the magnitude of this effect remains relatively constant throughout learning. Given the natural prevalence of the gambler's fallacy (i.e. negative recency) in sequential decision making tasks, it's possible that some kind of preexisting biases influenced their performance in the task (Gilovich, Vallone, & Tversky, 1985; Jarvik, 1951; Nicks, 1959). LASR also shows the learning effect in Figure 2 (top panel) in the first block of learning due to it's rapid adaptation to the task. However, assuming Boyer, et al.'s interpretation of the human data is correct (and not a floor effect of RT as participants gain experience in the task), it would be straightforward to simulate an initial bias in LASR by initializing the learning weights with a slightly negative value instead of zero at the beginning of learning.

Statistical Word Learning

It is clear from the previous simulations that despite it's simplicity, LASR can provide an accurate description of sequential learning behavior in the SRT. However, a key question remains concerning the generality of these findings: is sequential learning in the SRT sub-served by similar mechanisms as other areas of cognitive processing which rely on sequence processing? To evaluate this hypothesis, we apply LASR to the infant word learning study conducted by Saffran, Aslin, and Newport (1996).

Saffran et al. (1996) familiarized 8-month-old infants with a 2-minute recording of a computer-synthesized voice evenly reading a continuous stream of syllables at an even tempo. The stream was composed of four three-syllable nonsense words which were repeated in random order (examples word are "tu-pi-ro" and "go-la-bu"). The only cues concerning the beginning and end of words in the stream was the transitional probabilities between syllables which were higher between two syllables which occurred together within a words than between two syllables which spanned word boundaries.

On each trial of the test phase, infants were presented with repetitions of one of four three-syllable test strings. In Experiment 1, two of the test words were the same nonsense words which were presented during the familiarization phase while the remaining two were three syllable *non-words* which contained the same syllables heard during the familiarization phase but in a different order than they appeared in the initial phase. In Experiment 2, the test phase contrasted knowledge about *words* versus *part-words* where *part-words* consisted of syllables arranged in the same order as during familiarization, without directly corresponding to any of the words used to generate the familiarization sequence.

The results of both studies are shown in Figure 4 and indicate that infants were able to discriminate words from both non-words (Experiment 1) and part-words (Experiment 2) as reflected by longer listening times for the latter test stimuli. These findings demonstrate that infants are able to extract information about the statistical properties of a sequence given even a short incidental exposure to auditory stimuli.

Simulation Results To simulate these results with LASR, each syllable was treated as a separate event in the model. In both Experiment 1 and 2 there were 12 possible syllables, thus the model had 12 detectors. On each simulated trial, the model attempted to predict the next syllable in the sequence given the syllables which it had experienced so far.

During each trial of the the test phase, the memory register was cleared by setting all values back to zero and the output of the correct detector was recorded following the presentation of each syllable of the test sequence. In order to compare infant looking time and model performance (a necessarily indirect relationship), the output o_k of the correct detector for each syllable of the test sequence was summed to compute an overall familiarity score for the test item. These familiarity scores were then related to looking time via linear regression.

Figure 4 shows resulting performance of the model averaged over 1000 simulated experiments. The model predicts increased looking time for both non-words and part-words. Examination of the final setting of the detector weights reveal that the weights grow to approximate the transitional probabilities between syllables at different lags in the training sequence.

LASR provides a similar account of sequence learning in both the Lee (1997) and Saffran, et al. (1996) experiments. In each case, the model's weight grow to approximate the lag-n transition probabilities in the sequence (i.e. the probability of an event at time t given

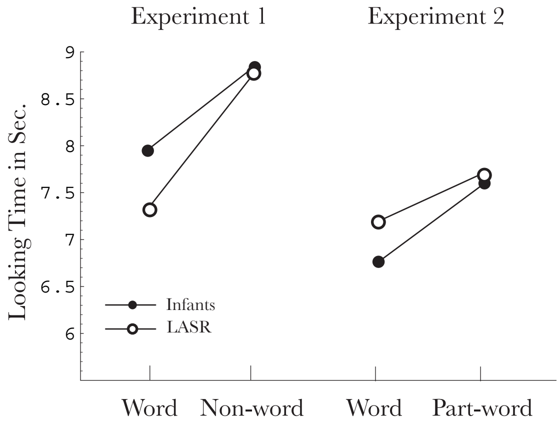

Figure 4: Comparison of infant and LASR results for Saffran, et al. (1996) Experiment 1 and 2.

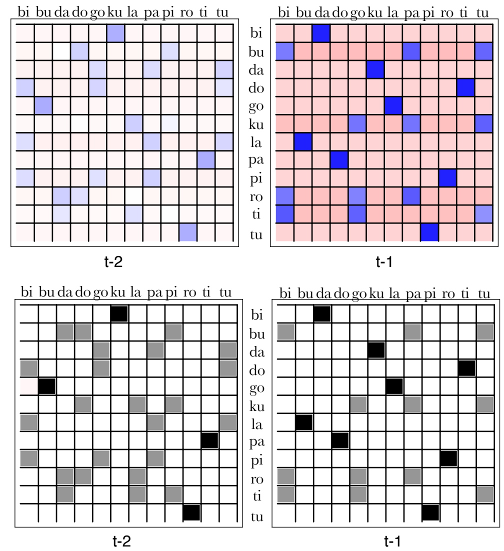

Figure 5: The final LASR weights for the Saffran, et al. (1996) infant word learning experiment. The labels on the rows correspond to possible syllable outcomes at each time slot, while the labels on the columns refer to the corresponding detector. The bottom two matrices show the actual transition probabilities between syllables at lag-1 and lag-2 in the training sequence. Black, grey, and white squares represent a transitional probability of 1.0, 0.33, and 0.0, respectively. The model weights closely mirror the transition probabilities.

a particular event on on trial $t - n$, Remillard & Clark, 2001). This is made clear in Figure 5 which compares the final setting of LASR's weights (top) and the true transition probabilities in the training sequence (bottom) at lag $t - 1$ and $t - 2$. LASR natural picks up on the statistical structure of the sequence and allows it to extract what might appear to be segmented knowledge about the sequence. The SRN has also been used to explain sequential word learning results similar to those studied by Saffran, et al. (Elman, 1990; Allen & Christiansen, 1996). However, the type of information acquired by the SRN differs from LASR because the SRN is capable of learning the true second order conditional probabilities due to it's hidden unit representations. The testing procedure used with infants does not distinguish between these two types of learning, but, (as these simulations show) the simpler lag-n statistic is sufficient to account for learning.

Conclusions and Discussion

The results of these simulations offer two conclusions. First, we show how the types of sequential learning reported in both the SRT and statistical word learning paradigms might be accounted by the same simple principles which define the LASR model. Recently, a number of authors have argued that behavior in both types of tasks could tap similar learning processes. Evidence in support of this hypothesis includes the fact that the type of sequential learning demonstrated by infants with artificial syllable languages has replicated to more general auditory stimuli such as tones (Saffran, Johnson, Aslin, & Newport, 1999) and to motor sequences in the SRT task (Hunt & Aslin, 2001) suggesting that this type of processing is not specific to linguistic material. Cross-species comparisons show that non-human primates are also able to discriminate words and non-words in the syllable task, again in support of the idea that learning in such tasks is not a property of a language specific learning system (Conway & Christiansen, 2001).

Second, we showed how a simple, single layer learning mechanism is able to account for findings which have previously been accounted for using more complex mechanisms. A full evaluation of LASR is not possible in this short paper, but preliminary work suggests that the model provides a similar account of the processes underlying implicit learning in many other studies. With this in mind, we offer LASR model as a possible "null" model for implicit sequence learning studies. It is important when developing models based on indirect measures of knowledge (such as reaction time) that theories aren't developed which reach beyond the data. We argue that LASR provides a tight match between the complexity of the model and the demonstrated processing complexity of the learner. In this sense, our argument bears some resemblance to other arguments put forward in the SRT literature (Perruchet, Gallego, & Savy, 1990; Reed & Johnson, 1994; Remillard & Clark, 2001). However, we build upon these criticisms by providing a viable modeling framework which shows promise as both an explanation and as a tool.

Acknowledgments Many thanks to Matt Jones, David Gliden, Levi Larkey, and Yasu Sakamoto for helpful many helpful conversations during the development of these ideas. This work was supported by Grant FA9550-

04-1-0226 and National Science Foundation CAREER Grant 0349101 to Bradley C. Love.

References

Allen, J., & Christiansen, M. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the eighteenth annual conference of the cognitive science society* (p. 370-375). Mahwah, NJ: Lawrence Erlbaum Associates.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (1998). The serial reaction time task: Learning without knowing, or knowing without learning? In *Proceedings of the 20th annual meeting of the cognitive science society* (p. 167-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing.* Cambridge, MA: MIT Press.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Science*, *2*(10), 373-417.

Cleeremans, A., & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235-253.

Cohen, A., Ivry, R., & Keele, S. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16*(1), 17-30.

Conway, C., & Christiansen, M. (2001). Sequential learning in non-human primates. *Trends in Cognitive Science*, *5*(12), 539-546.

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

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295-314.

Hunt, R., & Aslin, R. (2001). Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*(4), 658-680.

Jarvik, M. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology*, *41*, 291-297.

Keele, S., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2004). The cognitive and neural architecture of sequence representation. *Psychological Review*, *110*(2), 316-339.

Lebiere, C., & Wallach, D. (2000). Sequence learning in the act-r cognitive architecture: Empirical analysis of a hybrid model. In R. Sun & C. Giles (Eds.), *Sequence learning* (p. 188-212). Berlin: Springer-Verlag.

Lee, Y. (1997). Learning and awareness in the serial reaction time. In (p. 119-124). Hillsdale, NJ: Lawrence Erlbaum Associates.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Westport, Conn.: Greenwood Press.

Nicks, D. (1959). Prediction of sequential two-choice decisions from event runs. *Journal of Experimental Psychology*, *57*(2), 105-114.

Perruchet, P., Gallego, J., & Savy, I. (1990). A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology*, *22*, 493-516.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*(3), 585-594.

Remillard, G., & Clark, J. (2001). Implicit learning of first-, second-, and third-order transitional probabilities. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*(2), 483-498.

Rescorla, R., & Wagner, A. (1972). A theory of pavolvian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (p. 64-99). New York: Appleton-Century-Crofts.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386-408.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing. explorations in the microstructure of cognition.* Cambridge, MA, USA: MIT Press.

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

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

Sejnowski, T., & Rosenberg, C. (1987). Parallel networks that learn to produce english text. *Complex Systems*, *11*, 145-168.

Sun, R., Sluzarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*(1), 159-192.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention Record*, *4*, 96-104.