The Interaction of Perception and Cognition: A Competitive Connectionist Model of the Effects of Experience on Perceptual Representations

CMU-RI-TR-99-24

Lisa Marie Saksida

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Robotics Institute and Center for the Neural Basis of Cognition

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA

August, 1999

Thesis Committee:
Jay McClelland (Chair)
Tai Sing Lee
Tom Mitchell
John Pearce (Cardiff University, UK)

Copyright © 1999, Lisa Marie Saksida

ABSTRACT

Psychological models traditionally separate brain function into three stages of processing: perception, cognition, and action. While this compartmentalization strategy has been used effectively to simplify and thus better understand mental processing, a growing body of evidence for the phenomena of perceptual learning makes it clear than perception and cognition are inextricably linked (Hall, 1991). Thus, it seems that cognitive processes do not operate on static perceptual building blocks, but instead stimulus representations are flexible and adaptable to the task at hand.

In this dissertation, I explore the psychological and behavioral phenomena of perceptual learning through the construction of a model based on competitive learning, a standard algorithm used in the field of machine learning. The model provides a mechanism for nonassociative differentiation (Gibson & Gibson, 1955) and, in contrast to other models (e.g.,McLaren, Kaye, & Mackintosh, 1989), is compatible with a configural model of associative learning. The present model can account for critical perceptual learning phenomena such as exposure learning and effects of similarity on discrimination. It is also shown that the model can explain the paradoxical result that preexposure to stimuli can either facilitate or impair subsequent discrimination learning. In the dissertation, I have also provided an analysis of learning phenomena that are normally explained as being a result of "dimensional attention", and have shown that many of these phenomena may be more parsimoniously explained by the aforementioned core model of differentiation. Furthermore, I have extended the core model to show how the integration of reinforcement and categorization information with the stimulus representation may contribute to the phenomena of acquired distinctiveness and equivalence. Finally, I have also extended the core model to address issues regarding object representation in the brain; specifically, I have developed a model of the effects of lesions of monkey perirhinal cortex.

ACKNOWLEDGEMENTS

First, I would like to thank my dissertation advisor, Jay McClelland, for offering me the perfect balance between independence and guidance, for being quick to interpret the garbled expression of my ideas and then helping me turn them into feasible models, and for being wonderfully supportive throughout the dissertation process.

I am also very grateful to …

John Pearce for inspiring conversations about, and insightful comments on, this work;

the other members of my thesis committee, Tom Mitchell and Tai Sing Lee, for very constructive comments;

Dave Touretzky for his support of my interdisciplinary aspirations and for showing me how computer scientists think;

Barb Dorney, Jackie Jenkins, and Rebecca Sciullo for a refreshing dose of humanity and for helping to make the CNBC a fun place to be;

and all of the others at Carnegie Mellon who inspired and advised.

Other thanks in no particular order

Thanks to my parents for their constant faith even when my path may have seemed rather odd; and to my brother, Greg, for always managing to put things into perspective. Thanks to Lin Chase for taking me under her wing, showing me the ropes, and providing both personal and professional inspiration. Thanks to Kyra Straussman for her articulate company on long runs, for her life of drama, and for her general wisdom. Thanks to Belinda Thom for understanding. More thanks to Belinda, and to Robbie Warner, for making sure that I always had a place to stay and for being the best hosts imaginable. Thanks to Joseph, Boris, and Jim for always being ready for Mardi Gras, existential discussions, or at least driving by Pegasus. Thanks to Milky for her exemplary attitude toward life.

And, of course, thanks to Tim Bussey, my partner and colleague, for his reading of countless drafts of my work, for his contagious scientific enthusiasm, and for his unwavering support in the face of tremendous obstacles.

This work was supported in part by National Science Foundation grant IRI-9530975.

TABLE OF CONTENTS

Abstract ..... **ii**

Acknowledgements ..... **iii**

Table of Contents ..... **iv**

Chapter 1 Introduction **1**

Motivation 1
Approach 2
Outline 3

Chapter 2 Models of Learning **5**

The Associationist Approach 5
The Rescorla-Wagner Model *6*
Application to Discrimination and Perceptual Learning *8*
The Connectionist Approach 9
Learning in an Artificial Neural Network *10*
Supervised and Unsupervised Learning *10*
Supervised Learning and Backpropagation of Errors 11
Unsupervised Learning 14
Summary 16

Chapter 3 Exposure Learning **17**

Introduction 17
A Nonassociative Model of Perceptual Learning 20
Nonassociative Preprocessor *22*
Associative Processor *28*
Interaction between the Nonassociative and the Associative Processor *29*
Behavior *30*

Summary *31*
General Methods 31
Experiment 1: Generalization 32
Methods *32*
Results and Discussion *33*
Experiment 2: Peak Shift 34
Methods *35*
Results and Discussion *35*
Experiment 3: Similarity 37
Methods *38*
Results and Discussion *38*
Experiment 4: Preexposure facilitates a difficult discrimination 39
Methods *39*
Results and Discussion *39*
Experiment 5: Preexposure impairs an easy discrimination 41
Methods *41*
Results and Discussion *42*
Experiment 6: Separable latent inhibition and differentiation 43
Methods *43*
Results and Discussion *44*
Experiment 7: Simultaneous latent inhibition and perceptual learning 46
Methods *47*
Results and Discussion *47*
General Discussion 49
Summary *51*

Chapter 4 Dimensional Attention 52

Introduction 52
Evidence for Attentional Processing in Discrimination Learning *53*
Theoretical Interpretations *55*
Analyzer Theory 56
ALCOVE 57
Accounting for Dimensional Attention Effects without Attention 59
General Methods 60
Experiment 1: Overtraining Reversal Effect (ORE) 61
Methods *61*
Results and Discussion *62*
Experiment 2: Transfer of Learning Across a Continuum 63

Methods *64*
Results and Discussion *64*
Experiment 3: Further Analysis of Transfer Along a Continuum 66
Methods *67*
Results and Discussion *68*
Experiment 4: Comparison of Similar Stimuli 70
Methods *71*
Results and Discussion *71*
Experiment 5: Comparison of Easily Discriminable Stimuli 74
Methods *75*
Results and Discussion *76*
Experiment 6: Further Analysis of Stimulus Comparison 77
Methods *78*
Results and Discussion *79*
Experiment 7: Extradimensional versus Intradimensional Shifts 81
Methods *81*
Results and Discussion *82*
General Discussion 87
The ED/ID effect *87*
Real-Time Effects *88*
Summary *90*

Chapter 5 Acquired Equivalence and Acquired Distinctiveness 91

Introduction 91
Evidence for Acquired Equivalence and Distinctiveness *92*
Theoretical Interpretations *96*
Adding Task-Specific Information to the Model 98
Architecture of the Extended Model *99*
Operation of the Extended Model *100*
General Methods 103
Experiment 1: Labeling Stimuli 104
Methods *105*
Results and Discussion *105*
Experiment 2: Differential Outcomes Effect 106
Methods *107*
Results and Discussion *107*
Experiment 3: The Influence of Categorization on Perception 108
Methods *109*

Results and Discussion *110*
Acquired Distinctiveness 110
Acquired Equivalence of the Irrelevant Dimension 111
Acquired Equivalence within the Relevant Dimension 112
Local Sensitization of a Dimension 114
General Discussion 115

Chapter 6 Application to Neuropsychology: Effects of Perirhinal Cortex lesions on Stimulus Representations 118

Introduction 118
Functional and Organizational Principles 120
Architecture of the Model 123
General Methods 125
Experiment 1: Stimulus Set Size Effects in Concurrent Discrimination Learning 125
Methods *126*
Results and Discussion *126*
Experiment 2: Configural Learning 129
Methods *130*
Results and Discussion *131*
Experiment 3: Greater Effects of Lesions in PRh on Retention versus Acquisition 133
Methods *133*
Results and Discussion *134*
Experiment 4: Memory Capacity versus Visual Information Processing Accounts of Lesion-Induced Deficits 137
Methods *138*
Results and Discussion *139*
General Discussion 141
The Adaptive Value of Distributed Stimulus Representations *142*
Visual Information Processing versus Mnemonic Accounts of PRh Function *143*
A Note on Delayed Matching and Nonmatching-To-Sample *145*
Conclusion *147*

Chapter 7 Conclusions 149

Contributions 149
Further Questions 150

Chapter 8 References 153

Appendix A Psychological Models of Perceptual Learning 169

The associative view: McLaren, Kaye, and Mackintosh (1989) 169
The non-associative approach: Gibson and Gibson (1955) 171

Appendix B **Perirhinal Cortex: A Region at the Interface between Perception and Cognition** **173**

CHAPTER 1 INTRODUCTION

Motivation

Theoretical models of cognition usually address processes that are said to occur after the perceptual system has completed its work. According to such views, output from the perceptual system merely provides the initial information that is fodder for subsequent cognitive operations. This approach is common in the human cognitive psychology literature, in which a set of primitive elements usually forms the building blocks of cognition (e.g., Biederman, 1987; Julesz, 1981; Treisman & Gelade, 1980). Many researchers in associative animal learning theory have adopted a similar paradigm by assuming that any event that can be learned about will correspond to a static perceptual representation that is always fully activated in the presence of the stimulus. This division of behavior into serial stages reflects the sense-think-act paradigm that is also common in the fields of artificial intelligence and robotics. The strength of this approach is that it separates the processes contributing to behavior, thereby parsing a complex system into a form that may be more easily understood.

The assumptions upon which the above approach is based hold if perceptual representations are static. If perceptual organization changes as a result of experience, however, this should have a direct impact on cognitive processes. For example, the efficacy of a stimulus could be increased or decreased over time. The growing evidence for perceptual learning (PL) suggests that such changes can, indeed, occur and that perception may not be as stable as it is often assumed to be.

An alternative to the sense-think-act approach is to assume that perceptual building blocks are not fixed or finite but adapt to the requirements of the task for which they are employed. E. Gibson (Gibson, 1969), an early proponent of this

view, suggested that the perceptual interpretation of an entity depends on the observer's history, training, and acculturation. No set of primitives exists because the perceptual building blocks themselves are adaptive. Rather than providing a permanent concrete foundation, perception may instead support cognition by flexibly adapting to the requirements of cognitive tasks. So although cognitive processes involved in discrimination, attention, or object recognition may alter perception, the alteration is beneficial because as a result perception becomes better tuned to the task at hand.

The evidence for perceptual learning spans across several literatures including animal learning theory, cognitive and developmental psychology, and neuropsychology. These separate literatures often discuss, at different levels, what appears to be a similar process. The predominant goal of the current research, therefore, has been to construct a theoretical framework for the description and investigation of perceptual learning. This framework has been specified such that it can map to the different levels at which phenomena are discussed across the fields mentioned above.

Approach

Neural network, connectionist, or parallel distributed processing (PDP) models provide a bridge for relating cognitive processes directly to the underlying neural mechanisms. The PDP approach—based on the idea that cognitive processing arises from the interactions of large numbers of simple processing elements—was derived from the observation that the brain consists of just such a parallel processing system. This similarity in basic structure between the model and the brain facilitates the mapping of cognitive to biological processes. In addition, simulation modeling and mathematical analysis forces the clarification of postulated mechanisms and the assumptions upon which they depend. Furthermore, PDP models provide a method for making predictions that explicitly and demonstrably follow from a given psychological theory.

In this dissertation I have constructed a PDP model of phenomena associated with perceptual learning in order to provide a concrete framework that may help to synthesize data from across fields. In addition to providing necessary structure, this type of model fits well with extant theories of learning and cognition that focus on associative learning processes (e.g.,Pearce, 1994; Rescorla & Wagner, 1972).

Outline

The following chapter, Chapter 2, presents an overview of two approaches that have been taken toward understanding the phenomena of learning and memory: the associationist approach that dominates the study of animal learning and the connectionist approach often taken in the field of machine learning.

In Chapter 3 I present a connectionist model that performs bottom-up perceptual learning. With this model, I provide a concrete mechanism for a non-associative theory of bottom-up perceptual learning called "differentiation" (Gibson & Gibson, 1955) and integrate it with an associative learning framework. A set of simulations demonstrate that the present model can account for critical PL phenomena such as exposure learning and effects of similarity on discrimination. It is also shown that the model can explain the paradoxical result that preexposure to stimuli can either facilitate or impair subsequent discrimination learning. Predictions made by the model are discussed in relation to extant theories of PL. Chapters 4, 5, and 6 are dependent on the description of the model in this chapter.

In Chapter 4 the core model presented in Chapter 3 is developed further, and it is shown that several phenomena which have usually been attributed to relatively high-level processes involving "dimensional attention" (Sutherland & Mackintosh, 1971) may be more parsimoniously explained as a result of simple differentiation.

Chapter 5 provides a discussion of top-down influences on perceptual learning, and how they may be incorporated into the core differentiation model. The model is extended such that representations of external influences such as reinforcement or

categorization information may be combined with stimulus representations. It is shown that this extension allows the model to account for effects such as acquired equivalence and distinctiveness.

In Chapter 6 the architecture of the model, but none of the underlying principles, is changed in order more closely to relate the psychological model of perceptual learning to the organization of representations in the brain. In particular, it is shown that a puzzling pattern of effects seen after lesions of perirhinal cortex (PRh) in monkeys can be accounted for in the current simple model if it is assumed that bottom-up perceptual learning and discrimination processes are the mechanisms underlying "object identification".

Chapter 7 summarizes the dissertation and discusses its implications as well as future directions.

Three appendices provide additional information which may be of interest only to specific audiences. Appendix A presents an overview of two major theories of perceptual learning: nonassociative differentiation as described by Gibson and Gibson (1955) and the elemental, associative theory of McLaren, Kaye, and Mackintosh (1989). While not critical to an understanding of the current model, this information may be useful for understanding historical and theoretical arguments made in Chapter 3. Appendix B provides a brief overview of the anatomy of the ventral visual stream and the medial temporal lobe of the macaque monkey brain, and outlines the arguments for the involvement of PRh in memory and in perception.

CHAPTER 2 MODELS OF LEARNING

For many years, psychologists have been developing models of the ways in which animals learn in an effort better to understand how the human brain works. Machine learning theorists, on the other hand, have been developing learning models with the goal of building intelligent agents to help us with our daily tasks. Although the end goals pursued by these two groups of researchers are different, the more intermediate goal of understanding the learning process is the same. Nonetheless, the specific paths that have been taken toward this goal are understandably quite different as a result of the specific biases intrinsic to the different fields. A cross-fertilization of the ideas coming out of these different approaches should be beneficial for both fields in that insights gained in one field should be applicable to the other. This is the major goal of this dissertation: to apply an artificial neural network technique called competitive learning to several phenomena in the animal learning literature in an attempt to develop a model of the effects of experience on stimulus representations. In this chapter, I provide a brief introduction to the approaches used to develop learning models in these two fields.

The Associationist Approach

Most theories of the mechanisms underlying animal learning have been situated within the associative framework. The concept at the center of this associative framework is that central, or cognitive, representations of specified elements become linked so that activation of one can lead to the activation of its associate. The assumptions underlying the "standard associative model" (Roitblat, 1987) might be summarized as follows:

- elements are central representations of environmental events (stimuli), features of the stimuli, or responses.

- elements are directly activated by the presentation of the appropriate stimulus or feature.

- associations are directional connections between pairs of elements.

- an associative connection allows activity in one element to modify the state of activation of another.

- all learning consists of the strengthening of such associative connections.

The Rescorla-Wagner Model

A recent, very influential associative model of learning is that of Rescorla and Wagner (1972). The model provides a formal specification of the conditions under which associations are formed and strengthened. The central concept of the model is that of associative strength (V), which is the strength of a directional connection between elements representing the two stimuli used in studies of Pavlovian or classical conditioning. During a Pavlovian conditioning trial, the first stimulus that is presented is called the conditioned stimulus (CS) and usually consists of a neutral event such as a light or a tone. The second event, which is typically presented at the offset of the CS, is referred to as the reinforcer, outcome, or unconditioned stimulus (US). The US is usually a stimulus of motivational significance (e.g., shock or food). The associative strength of the CS-US connection increases when the two events occur contiguously, that is, when the central representations of the two events are concurrently active. A conditioned response (CR) consists of a learned response to the initially neutral CS. The likelihood of a CR is assumed to increase with V, and such a response is used as an index of V in the animal.

According to the Rescorla-Wagner rule, the change in associative strength produced by a conditioning trial is:

$$\Delta V_x = \alpha_x \beta \left(\lambda - \sum_x V_x \right) \quad (2\text{-}1)$$

where V_x is the associative strength on the link between element x and the US, a_x is the associability (learning rate) of element x and depends on the intensity, discriminability, or salience of x, β is the associability of the reinforcer and depends on reinforcer intensity, and λ is the maximum associative strength that the US can support.

Conditioning with a single CS proceeds as follows. Over successive trials there are increments in V_A until the point is reached at which $(\lambda - \sum V) = 0$ and no further increases are possible. Thus, this account of conditioning suggests that the US grows less and less effective as conditioning proceeds. When two or more CSs occur together, Rescorla and Wagner assume that each stimulus acts independently, thus the associative strength of a light and tone presented together would consist of the sum of the separate strengths of the stimuli. In addition to excitatory associations, inhibitory associations can also develop. For example, take the situation in which a subject is trained to asymptote with one CS, A, and is then presented with trials of the compound AB without reinforcement. λ is 0 when no US is presented, so the change in associative strength is negative on these trials and B develops an inhibitory link to the outcome representation. Thus, presentation of B in future can inhibit the activation of the US representation.

The Rescorla-Wagner model is a prime example of Roitblat's standard associative theory. The elements of association are stimuli, associations are created as a result of contiguous stimulus presentation, and there are two kinds of learning (excitatory and inhibitory association). The popularity of this model stems largely from the fact that despite its remarkable simplicity, it can explain a large range of animal learning phenomena. One example is the way in which the model can account for blocking (Kamin, 1968). In the blocking paradigm, subjects are conditioned to asymptote with CS A. Then, in a second phase of training, they are

conditioned to the compound stimulus AB. In a subsequent test phase, subjects are presented with B alone. Performance on the test phase usually indicates that B has gained little associative strength as it only produces a weak CR. In controls which have only been trained on the second phase, however, B gains considerable associative strength. Thus pretraining with A is said to "block" subsequent conditioning to B when B is presented in compound with A. This effect could be attention-based: pretraining with A causes the animal to attend more to it therefore it ignores B during the second training phase. But the Rescorla-Wagner allows for an alternate explanation that is simpler: the presence of A ensures that the value of $\left(\lambda - \sum V\right)$ equals zero on compound trials therefore the reinforcer is ineffective and the associative strength of B remains at zero.

Application to Discrimination and Perceptual Learning

In this dissertation, I am mostly concerned with discrimination learning rather than simple conditioning. During discrimination training, the experimenter presents two CSs, either serially or simultaneously, that differ in reinforcement contingency and subjects learning to respond only to the positive CS. There are two polarized ways of thinking about how this type of learning works: (1) discrimination learning is solely about how animals learn to respond differentially to the two stimuli and (2) the stimuli were perceived as identical at first and one consequence of discrimination training is to make them perceptually distinguishable (e.g., Konorski, 1967). The former idea, that discrimination learning is only an associative process, is the explanation provided by the standard associative model. The latter idea, that perceptual representations are altered during discrimination training, has not received as much attention because of the focus in the literature on the associative model. This idea of perceptual learning—whether it exists and how we may account for it—is the topic explored in detail in the remainder of this thesis.

The Connectionist Approach

The connectionist approach to learning is highly compatible with the associationist approach described above. An artificial neural network consists of a set of model neurons with connections between them. McCulloch and Pitts (1943) proposed a simple model of a neuron as a binary threshold unit. The model neuron computes the weighted sum of its inputs from other units and outputs a one or a zero depending on whether the sum is less than or greater than the threshold. Figure 2-1 is a schematic diagram of the neuron and Equation 2-2 shows how the activation of a neuron is determined for the next time step.

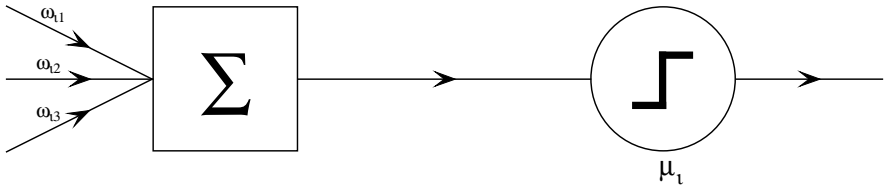


Figure 2-1: The McCulloch-Pitts neuron. After Hertz, Krogh and Palmer (1991, p. 3).

$$x_i(t+1) = \Theta(\sum_j w_{ij} x_{ij}(t) - \mu_i) \tag{2-2}$$

where x_i equals 1 or 0 and represents the activation of neuron *i*, time *(t)* is discrete and one time unit elapses for each processing step, $\Theta(x)$ is the step or Heaviside function

$$\Theta(x) = 1 \text{ if } x \geq 0; 0 \text{ otherwise} \tag{2-3}$$

w_{ij} is the synapse connecting neuron *i* to neuron *j*, and μ is the threshold value for neuron *i*.

In most artificial neural networks units something like the McCulloch-Pitts neuron are arranged in layers with full forward connectivity from the previous layer (i.e., each unit in the first layer is connected to every unit in the second layer). Units

in the first layer have no incoming weights; instead they are clamped at a certain level of activation by an input pattern. The final layer contains output units and the network may also contain "hidden" units so named because they connect to the output units but are not directly observable (see Figure 2-2).

Learning in an Artificial Neural Network

Artificial neural networks compute a function in which each input pattern is mapped to an pattern of activation at the output units. The weights on the connections between units can be altered and so change the function that is computed.

The rule used to change connection weights is usually a modification of Hebb's (1949) rule formulated to explain learning in real neurons: the connection between the pre- and postsynaptic cell is strengthened when both cells are active simultaneously. As such, the learning rule used in many artificial neural networks such as the associative memory model (see Hinton & Anderson, 1989) is a function of the degree of activity of units on both sides of the connection:

$$\Delta w_{ij} = \eta x_i x_j \tag{2-4}$$

where w_{ij} is the weight on the connection between unit i and unit j, η is a parameter that controls the degree of weight change on a trial, x_i is the activation of unit i and x_j is the activation of unit j.

Supervised and Unsupervised Learning

One dimension along which artificial neural networks may be categorized is that of degree of supervision. In supervised learning extra information about the desired output is given to the network in the form of either the actual desired output or some type of reinforcement signal. In unsupervised learning, on the other hand, only

input patterns are received and the algorithm tries to find patterns intrinsic to the structure of the input distribution.

Supervised Learning and Backpropagation of Errors

In most supervised learning algorithms, the goal is to minimize the difference between the actual and the desired outputs (the error). The usual mathematical form for this error is:

$$E(\mathbf{W}) = \frac{1}{2}\sum_{\xi}\sum_{i}\left(y_i^* - y_i^{\xi}\right)^2 \tag{2-5}$$

where y_i^* and y_i^{ξ} are the desired and actual outputs of the i^{th} neuron when the network is presented with pattern $\boldsymbol{\xi}$.

A common technique for minimizing the error is called gradient descent. This involves calculating the gradient of the error function with respect to the weights of the connections and taking a small step in the opposite direction; the basic idea is that in order to find the minimum point in the error function one should move downhill If the error function is given by $E(w_{ij})$, the gradient descent update (or delta) rule is

$$\Delta w_{i,j} = -\eta \frac{\partial E}{\partial w_{i,j}} = \eta(y_i^* - y_i^{\xi})\xi_k \tag{2-6}$$

where w_{ij} is the weight on the connection between unit i and unit j, η is a parameter that controls the degree of weight change on a trial, and y_i^* and y_i^{ξ} are the desired and actual outputs of the i^{th} neuron when the network is presented with pattern ξ_k.

The backpropagation algorithm (Rumelhart, Hinton, & WIlliams, 1986; Werbos, 1974) is central to the operation of multi-layer neural networks as it provides a method for applying gradient descent to connections throughout the network, not just the output unit connections, in order to learn a training set of input-output pairs. Consider the two-layer network shown in Figure 2-2.

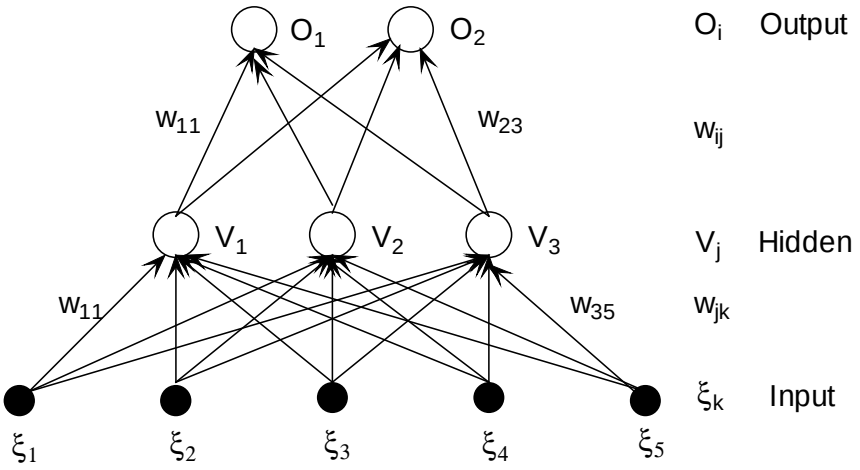


Figure 2-2: A two-layer feedforward artificial neural network. After Hertz, Krogh and Palmer (1991, p. 116).

For this two-layer network, output units are denoted by O_i, hidden units by V_j, and input units by ξ_k. There are connections w_{jk} from the inputs to the hidden units and W_{ij} from the hidden units to the output units.

Given an input pattern, hidden unit *j* receives a net input of

$$h_j = \sum_k w_{jk}\xi_k \tag{2-7}$$

and produces an output of

$$V_j = g(h_j) = g(\sum_k w_{jk}\xi_k). \tag{2-8}$$

Output unit *i* thus receives

$$h_i = \sum_j W_{ij}V_j = \sum_j W_{ij}g(\sum_k w_{jk}\xi_k) \tag{2-9}$$

and produces final output

$$O_i = g(h_i) = g(\sum_j W_{ij}V_j) = g(\sum_j W_{ij}g(\sum_k w_{jk}\xi_k)). \tag{2-10}$$

(Note that thresholds are omitted as they can be taken care of by an extra input unit clamped to –1 and connected to all units in the network.)

The usual error or cost function, specified in Equation (2-5), now becomes a continuous differentiable function of every weight so gradient descent can be used to minimize error

$$E(\mathbf{w}) = \frac{1}{2}\sum_i \left[y_i^* - g(\sum_j W_{ij} g(\sum_k w_{jk}\xi_k)) \right]^2 . \tag{2-11}$$

The resulting update rule for the hidden-to-output connections given by the gradient descent rule (Equation 2-6) is

$$\Delta w_{i,j} = -\eta \frac{\partial E}{\partial w_{i,j}} = \eta \left[y_i^* - y_i^{\xi} \right] g'(h_i) V_j = \eta \delta_i V_j \tag{2-12}$$

where we have defined

$$\delta_i = \left[y_i^* - y_i^{\xi} \right] g'(h_i) . \tag{2-13}$$

For the input-to-hidden connections one must differentiate with respect to the hidden-to-output connections. Using the chain rule

$$\begin{aligned} \Delta w_{jk} &= -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \frac{\partial E}{\partial V_j} \frac{\partial V_j}{\partial w_{jk}} \\ &= \eta \sum_i \left[y_i^* - y_i^{\xi} \right] g'(h_i) W_{ij} g'(h_j) \xi_k \\ &= \eta \sum_i \delta_i W_{ij} g'(h_j) \xi_k \\ &= \eta \delta_j \xi_k \end{aligned} \tag{2-14}$$

with

$$\delta_j = g'(h_j) \sum_i W_{ij} \delta_i . \tag{2-15}$$

Equation (2-15) allows one to determine the δ for a given hidden unit V_j in terms of the δ's of the units y_i that feeds. The coefficients are the usual W_{ij}'s, but here they propagate errors (δ) backwards instead of activation signals forward.

Unsupervised Learning

A basic form of unsupervised learning is called competitive learning (Grossberg, 1976a; Grossberg, 1976b; Kohonen, 1982; Rumelhart & Zipser, 1986). In competitive learning, only one output unit is allowed to fire at a time. The output units compete to be the unit that fires, and are thus called winner-take-all units.

The object of competitive learning networks is to categorize the input data. The idea is that similar inputs should be classified as being in the same category and so should fire the same output unit. The classes must be found by the network itself—no desired output information is provided—from the correlations of the input data.

In the simplest competitive learning networks there is a single layer of output units O_i fully connected to a layer of inputs ξ_j via weighted connections w_{ij} (see Figure 2-3).

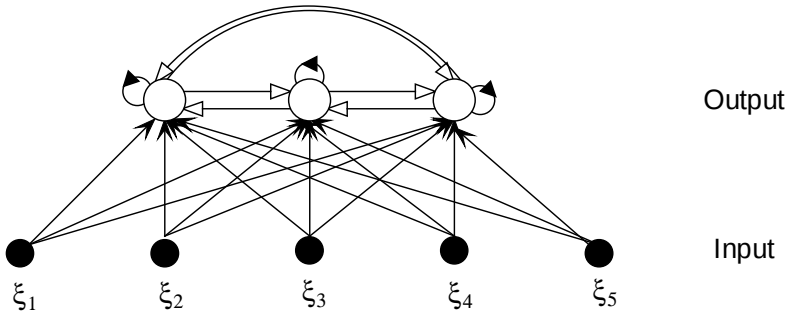


Figure 2-3: A simple competitive learning network. Open arrows denote inhibitory connections and closed arrows denote excitatory connections.

The winner is the usually the unit with the largest net input. The input for the current input vector $\boldsymbol{\xi}$ is

$$h_i = \sum_j w_{ij}\xi_j = \mathbf{w}_i \cdot \boldsymbol{\xi} \tag{2-16}$$

thus

$$\mathbf{w}_{i^*} \cdot \boldsymbol{\xi} \geq \mathbf{w}_i \cdot \boldsymbol{\xi} \text{ (for all } i\text{)} \tag{2-17}$$

defines the winning unit i^* with $y_{i^*}=1$.

In a computer simulation, the winner is normally determined by searching for the maximum h_i. In a real neural network, it is possible to implement this with lateral inhibition as shown in Figure 2-3. Each competitive unit inhibits the others, but also has a self-excitatory connection. The appropriate lateral weights and nonlinear activation function must be chosen well in order to avoid oscillation (see Grossberg, 1976a; Grossberg, 1976b).

In order for the network to learn to cluster the input patterns, it operates as follows. The weight vectors $\mathbf{w}_i$ are set to small random values. Then a set of input patterns $\boldsymbol{\xi}$ is presented to the network. For each input the winner i^* among the outputs is determined and then the weights w_{i^*j} for the winning unit only are updated in order to make the $\mathbf{w}_{i^*}$ vector a bit closer to the input vector. This makes the winner more likely to win when that input is presented in future. This is done using the following update rule

$$\Delta w_{i^*j} = \eta(\xi_j - w_{i^*j}) . \tag{2-18}$$

Kohonen Feature Mapping (Kohonen, 1982; Kohonen, 1984) is similar to competitive learning except that output neurons have a specified positional relationship to each other in output space. The network topology is usually a two-dimensional grid, motivated by the structure of the cortex. As in competitive learning, the weights of the output units define positions in the input space. Unlike competitive learning, however, not just the winning unit's weights, but also surrounding unit's weights, are updated on a cycle. The neighboring unit's weights

are updated with a distance dependent attenuated version of the weight change. The end result is that nearby units respond to similar patterns. The Kohonen weight update formula is

$$\Delta w_{ij} = \Lambda(j,j^*)\eta(\xi_i - w_{ij}) \text{ for all } j \tag{2-19}$$

where $\Lambda(j,j^*)$ is a monotonic decreasing function of $\left\| r_j - r_{j^*} \right\|$, j^* is the index of the winning unit and r_j gives the position of the unit with index j in the topology within its layer.

Summary

In this chapter I've provided the briefest of overviews of associative animal learning theory and connectionist neural network learning. My intent was simply to present a very basic foundation for the theoretical issues that will be discussed throughout this dissertation. I think it is clear that whereas animal learning theorists and artificial neural network theorists use different languages and have different goals, some of the approaches are very similar (e.g., the Rescorla-Wagner rule in animal learning theory is equivalent to the delta rule used for gradient descent in neural network theory). For more information on animal learning theory, any of the following would be a good place to start Pearce (1987a), Dickinson (1980), or Mackintosh (1974). For more information on artificial neural networks see Hertz (1991), McClelland (1986), or Rumelhart (1986).

CHAPTER 3 EXPOSURE LEARNING

Introduction

A large body of data provides evidence for the idea that animals' internal representations of stimuli can change simply as a result of experience (see Hall, 1991 for a review). Much of this evidence derives from two contrasting behavioral effects. The first effect, "latent inhibition", refers to the phenomenon whereby prior non-reinforced exposure to a stimulus retards subsequent conditioning to that stimulus (Lubow & Moore, 1959). The second effect is "perceptual learning", in which preexposure to stimuli may facilitate subsequent learning of a discrimination between them (Gibson, Walk, Pick, & Tighe, 1958). As a result, while the ease of a discrimination between completely novel stimuli may initially depend only on their similarity, as an animal gains experience with the stimuli the ease of discrimination, as manifested in the rate of learning of the discrimination, may either increase or decrease.

Many researchers have successfully integrated latent inhibition into an associative learning framework. For example, Pearce and Hall (1980) attribute latent inhibition to a reduction in associability, or learning rate, of the conditioned stimulus (CS). They suggest that as the outcome of stimulus presentation becomes well predicted, attention to the stimulus (as manifested in associability) decreases. Thus when an animal is exposed to a stimulus which is followed by a consistent outcome (e.g., nothing in the case of latent inhibition) attention to the stimulus decreases. This reduction in attention leads to a reduction in the associability of the stimulus, and the animal learns a conditioned response (CR) to that stimulus relatively slowly. Wagner (1978; 1981), on the other hand, suggests that a stimulus will be fully processed only when its occurrence is not predictable on the basis of other, usually contextual, cues. Others (e.g., Lubow, Schnur, & Rifkin, 1976) have suggested that latent inhibition is

due to the disruption of the formation of new associations through interference or competition with associations learned during preexposure. A fourth group of theories suggest that CS associability is a function of the extent to which the stimulus predicts a particular outcome (e.g., Frey & Sears, 1978; Mackintosh, 1975; Moore & Stickney, 1980; Schmajuk, Lam, & Gray, 1996; Schmajuk & Moore, 1989), whereas a fifth group assumes that CS preexposure causes a failure of performance at the time of retrieval (e.g., Bouton, 1993; Kasprow, Catterson, Schachtman, & Miller, 1982; Miller & Schachtman, 1985; Spear, 1981).

Perceptual learning, however, has proven to be somewhat more problematic than latent inhibition for associative learning theory. One contributing reason is that the results from perceptual learning experiments are paradoxical: in some cases, preexposure facilitates subsequent discrimination whereas in other cases preexposure impairs discrimination. Gibson and Walk (1956), in the original demonstration of perceptual learning, exposed rats to black metal triangles and circles attached to the walls of their home cages. After some time, control subjects with no exposure and preexposed subjects were trained on a food rewarded discrimination with triangles and circles as the relevant stimuli. The experimental group performed much better than the controls. Subsequent work has confirmed that preexposure can facilitate discrimination learning (see Bennett & Ellis, 1968; Forgus, 1956; Kawachi, 1965). However, other experiments using similar procedures have obtained different results (e.g., Chantrey, 1972; Gibson, 1969). Chantrey (1972), for example, using an imprinting procedure showed that domestic chicks given prior experience with a pair of colors were in some circumstances *less* able to learn the discrimination than were control subjects.

Results such as the above prompted investigators to explore the parameters required to obtain the facilitation seen by Gibson and Walk (1956). While the type of stimulus, duration of exposure, and age of the subjects do not seem to be highly relevant to the preexposure effect, the difficulty of the discrimination does seem to be an important factor. Oswalt (1972) used a range of stimuli and tested

discriminations of different difficulties: triangles vs. circles, objects with differently oriented stripes, and panels with upright or inverted U-shaped figures. Only the first discrimination—the most difficult—was facilitated by preexposure. Chamizo and Mackintosh (1989) provided a second piece of evidence suggesting that difficulty of discrimination is important in perceptual learning. When rats were tested in a Y-maze in which the different arms were distinguished only by the texture of their flooring, preexposure facilitated discrimination. However, in a maze in which both the floor texture and the wall color differentiated the arms, the preexposed animals were *slower* to acquire the subsequent discrimination. Since the second discrimination provided a greater number of highly salient cues than the original discrimination, thus making the discrimination easier, it can be seen as providing evidence for the idea that preexposure helps difficult discriminations but hinders relatively easy ones. More recently, Honey, Bateson, and Horn (1994), in a chick imprinting paradigm, have demonstrated a similar result.

Several theorists have attempted to explain perceptual learning using associative principles (e.g., Hall, 1991; McLaren et al., 1989). Associative theories propose that, in addition to associations between elements of different events, associations also form between elements of a single event such as the CS itself. Perceptual learning then consists of refinement of the stimulus representation through the operation of associative processes upon stimulus elements. Others have offered nonassociative accounts, suggesting that perceptual learning involves the gradual abstraction from a complete stimulus of features relevant to the discrimination (Gibson & Gibson, 1955).

Neither of the above approaches, however, has provided a fully satisfactory theory of exposure effects in discrimination learning. Whereas the associative approaches have provided concrete mechanisms for both latent inhibition and perceptual learning, the models either do not address stimulus similarity or are generally subject to Pearce's (1994) criticisms of the way in which elemental models

handle similarity of stimuli[1]. The nonassociative approach, while being intuitively plausible, has not been developed sufficiently to provide a concrete mechanism by which perceptual learning may operate, and furthermore does not begin to address the issue of latent inhibition.

As an alternative to the above approaches, I have developed a nonassociative model of perceptual learning that provides a simple account of the finding that discriminations are sometimes facilitated and sometimes retarded by exposure. After Channell and Hall (1981) the model starts from the assumption that two processes—one which decreases associability and one which increases discriminability—operate during preexposure to stimuli.

A Nonassociative Model of Perceptual Learning

As discussed above, there exist many well-formulated theories of latent inhibition. As a result, the current model focuses on the perceptual learning mechanism, and I have simply chosen one theory of latent inhibition, that of Pearce and Hall (1980), to adjust attention to stimuli. The contribution of the current model, therefore, is that it provides a concrete mechanism for Gibsonian differentiation, an idea that is intuitively appealing but to this point has been poorly specified. In addition, in contrast to associative approaches which have been based on elemental learning theory, the current nonassociative perceptual learning mechanism is compatible with a configural model of discrimination learning (Pearce, 1994).

Figure 3-1 illustrates the structure of the discrimination and perceptual learning network. Inputs to the network consist of points in a two-dimensional stimulus

[1] Pearce's (1994) main argument against elemental models is that they sometimes make the incorrect and paradoxical prediction that stimuli with common elements will be more easily discriminated than those without common elements. In other words, a manipulation which increases the similarity of two stimuli is predicted to also increase their ease of discriminability. See Experiment 3 for more detail, or Pearce (1994) for a full discussion of the issue.

space. Input of a stimulus to the network leads to a "percept" consisting of the activation of a two-dimensional layer of input units. These input units feed into a two-dimensional layer of competitive units, the activation of which comprises a "cognitive" representation of the stimulus.

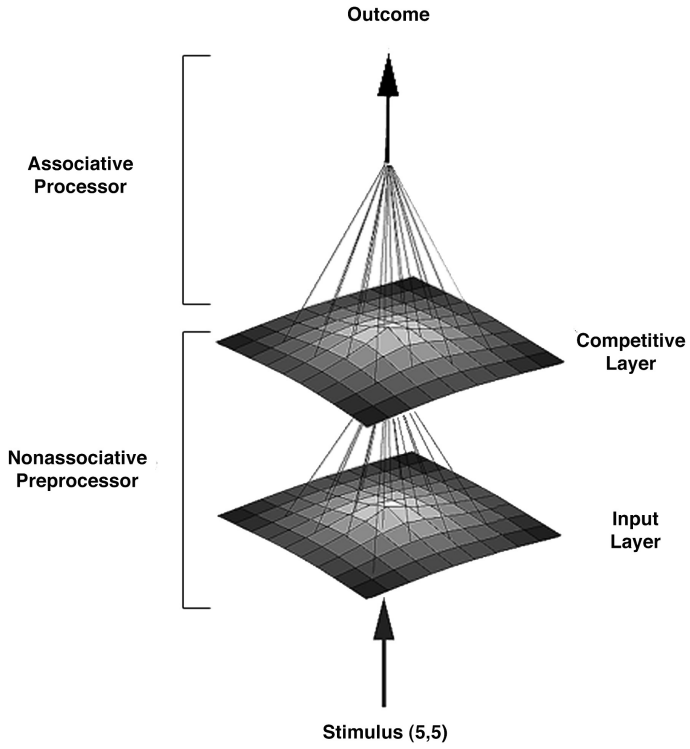


Figure 3-1: Overview of the perceptual learning network. The model consists of two layers of simple units: the input layer and the competitive layer. The input layer is fully connected to the competitive layer via a set of weights that are adjusted by a nonassociative (i.e., unsupervised) learning mechanism. The competitive layer is fully connected to a node representing the outcome of the presentation of a stimulus to the network (e.g., response or reinforcer). These weights are adjusted via an associative mechanism.

Perceptual learning occurs on the weights on the links between the input layer and each competitive unit. These weights are updated—proportionally to the competitive unit's activation—via a competitive learning mechanism that adjusts the weights of the competitive unit to be closer to the pattern of activation in the input

layer. Over time the weights of active competitive units become a more accurate representation of a repeatedly presented stimulus.

Each competitive unit maintains a parameter, α, which reflects its associative learning rate. The α values are adjusted on each trial in relation to both the unit's degree of activation and the discrepancy between the CR and reinforcement. This provides a straightforward mechanism for latent inhibition (Pearce & Hall, 1980). Competitive units also maintain a second parameter, σ, which reflects the extent to which a given competitive unit may affects its neighbors.

The competitive units are each linked to an outcome representation, and the weight on the link reflects the strength of association between a stimulus representation and a given outcome. When an input is presented, the competitive units are differentially activated according to their distance from the most activated competitive unit, and the sum of these activations multiplied by their associative strengths determines the strength of the CR. All associative weights are updated in proportion both to the degree of activation of the competitive unit and the discrepancy between the CR and the outcome.

The network can be thought of as consisting of two sections: (1) a nonassociative preprocessor comprising the mechanisms which operate on the weights between the input layer and the competitive layer and (2) an associative processor, comprising the mechanisms which operate on the weights between the competitive layer and the outcome representation. These are described in detail in the following sections. Note that vectors and matrices are formatted in bold type whereas elements of vectors and matrices are formatted in italics (e.g., a_j is the jth element of a matrix **a**).

Nonassociative Preprocessor

The nonassociative preprocessor consists of two layers of simple units: the input layer and the competitive layer. The input layer is fully connected to the competitive

layer via a set of weights that are adjusted by a nonassociative learning mechanism. Activation of units in the input layer represents initial processing of an external stimulus and is thus representative of a percept. Activation of units in the competitive layer, on the other hand, is reflective of the changeable representation of the stimulus.

A stimulus is represented as a pair of numbers, each of which represents the value of the stimulus on a dimension. In this chapter, stimuli are assumed to be bi-dimensional. Thus, for a red circular stimulus $\mathbf{S}=\langle x,y\rangle$, x might represent the color of the circle in nanometers and y might represent its diameter in centimeters.

On a training trial, the x and y values of a stimulus are fed into the network. This leads to a pattern of activation on the input layer **I** as follows. The input layer consists of a square set of simple units, with the range of the units being greater than the maximum x or y value of any stimulus that might be presented. Each unit i has a corresponding activation that ranges between 0 and 1. The activation of all units in the input layer is initially zero. When a stimulus $\mathbf{S}=\langle x,y\rangle$ is presented, however, the unit in the input layer whose coordinates correspond to the x and y values of **S** becomes maximally activated (i.e., its activation level is shifted from 0 to 1). This unit affects nearby units such that those that are within a certain radius (σ) of the unit will also be activated proportionally to their distance from it. This means that the degree of generalization between two stimuli can be deduced from either the amount of overlap in their respective input layer activations or by calculating the Euclidean distance between them. For example, two similar stimuli might be represented as $\mathbf{S}_1=\mathbf{<}2,2\mathbf{>}$ and $\mathbf{S}_2=\mathbf{<}3,5\mathbf{>}$ whereas two different stimuli would be represented as $\mathbf{S}_3=\mathbf{<}2,2\mathbf{>}$ and $\mathbf{S}_4=\mathbf{<}8,8\mathbf{>}$**.** Note that stimulus representations do not "wrap" around the edges of the grid, meaning that on a 10 by 10 unit grid, <10,10> is the furthest unit from <1,1>.

Equation 3-1 is the input layer activation function for a stimulus whose input layer location is $i^*=\langle x,y\rangle$. The degree of activation of a input layer element with

location i_j is 1 for $i_j=i^*$ and decreases as a function of the distance between i_j and i^*. The width of the Gaussian is σ^2.

$$a_j = \exp\left(-\frac{\left\|i_j - i^*\right\|^2}{2\sigma^2}\right) \tag{3-1}$$

Every unit in the input layer is connected to each unit in the subsequent competitive layer **C**. The link between one input unit *j* and one competitive unit *k* is called w_{jk} Thus for competitive unit *k* there exists a corresponding set of weighted links, $\mathbf{w}_k$, which consists of the set of links between each unit in the input layer and unit *k*.

In order that the competitive layer is roughly topographic from the start, w_{jk} are initialized as Gaussian distributions, centered over the corresponding input unit coordinates, that are corrupted by a significant degree of uniformly distributed random noise *r*.

$$w_{jk}(0) = \exp\left(-\frac{\left\|i_j - c_k\right\|^2}{2\sigma^2}\right) + r(-0.3,0.3) \tag{3-2}$$

In the above equation, i_j is an <*x,y*> location in the input layer and c_k is an <*x,y*> location in the competitive layer.

When a stimulus is presented to the network, the Euclidean distance between $\mathbf{w}_k$ and the pattern of activation on the input layer (**a**) are calculated (see Equation 3-4) and compared (see Equation 3-3) for each competitive unit.

$$k^* = \arg\min(\mathbf{d}) \tag{3-3}$$

$$d_k = \sqrt{\sum_j (a_j - w_{jk})^2} \tag{3-4}$$

As in the input layer, the winning unit influences neighboring units, and they are activated proportionally to their distance from the winner. Competitive unit activation a_k is 1 for $c_k = c^*$ and falls off with distance $\left\| c_k - c^* \right\|$ where c_k and c^* are the grid coordinates of units k and k^* respectively. As a result, units close to the winner are significantly more active than units further away. The degree of activation of a unit k is determined by a Gaussian defined as follows:

$$a_k = \exp\left(-\frac{\left\| c_k - c^* \right\|}{2\sigma^2} \right) \tag{3-5}$$

As mentioned above, the extent of influence, or neighborhood size, of the winning unit is reflected in the Gaussian width parameter σ.

The weights on the links between the input layer and the competitive layer are at the heart of the nonassociative learning mechanism that contributes to perceptual learning. Each time a stimulus is presented, the weights $\mathbf{w}^*$ of the winner are adjusted such that they are a bit closer to the pattern of activation on the input layer. Over time, the $\mathbf{w}^*$ become closer and closer to the input layer activation pattern **a**, thus the competitive unit weights of the winner become a better representation of the stimulus.

In addition to the winner, the $\mathbf{w}_k$ corresponding to a set of nearby units are also updated in proportion to their proximity to the winner. The result is that after training is complete, nearby competitive units respond to nearby input patterns. The effect of this is to set up regions of the competitive layer which code for similar representations, that is, those with common features. Equation 3-6 is responsible for the competitive weight update process.

$$\Delta w_{jk} = \delta_k \cdot a_k \cdot \left(a_j - w_{jk} \right) \tag{3-6}$$

The competitive learning rate, δ_k, decreases as the outcome of trials become well predicted (see Equation 3-10).

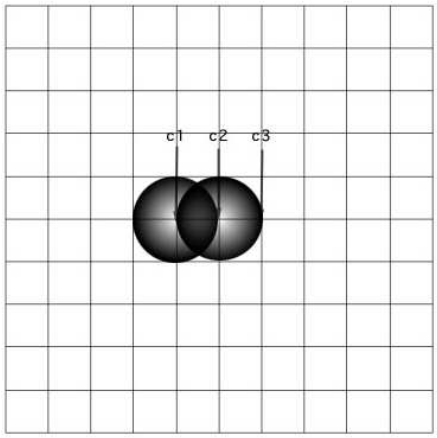
The nonassociative perceptual learning mechanism outlined above is based on competitive learning, a standard technique that is used for pattern classification in machine learning (Kohonen, 1984; Rumelhart & Zipser, 1986). It is "unsupervised" in that no external feedback to the system indicates whether or not a particular classification is appropriate. Instead, the algorithm exploits the statistical properties of stimuli and classifies them based on those properties. Over time, as the system gains more samples, the statistical classifications become more reflective of the parameters of the actual population being sampled. Thus, over time the mechanism causes the representation of the stimulus to become more accurate and robust.

As a result of this competitive mechanism, the discriminability of a stimulus tends to increase with exposure (perceptual learning). To illustrate, consider two similar input stimuli (e.g., S_1 and S_2). Such stimuli are usually captured by topographically close competitive units (e.g., c_1 captures S_1 and c_2 captures S_2), which means that the two winning units will tend to be in each other's neighborhoods (see Figure 3-2A). The closer that S_1 and S_2 are, the more activated c_1 will be when S_2 is shown. As a result, when S_1 and S_2 are repeatedly presented, a competition ensues in which each of the competitive units, when its corresponding input stimulus is delivered, moves its weight vector as well as the other winner's weight vector toward the value of the stimulus. That is, when S_1 is presented, c_1 moves both itself and c_2 toward S_1. On the next time step, when S_2 is presented, c_2 moves itself and c_1 toward S_2. While this is happening, the other competitive units which are in the winner's neighborhood are moved as well, proportionally to their distance from the winners. If they fall into both winner's neighborhoods, then they are moved back and forth between S_1 and S_2 as well. However, a unit c_3 which is in the neighborhood of c_2 but not in the neighborhood of c_1 will be affected only when c_2 wins. Over time, because c_2 is pulled toward S_1 but c_3 is not, c_3 develops a better representation of S_2 and wins over c_2 when S_2 is presented (see Figure 3-2B). This means that the new winning units, c_1 and c_3 are more separated on the competitive grid thus there is less overlap in their distributions of activation. Over time, as the above process continues, the two winning units become more and more separated on the competitive grid, thus

lessening their effect on each other and facilitating discrimination by lessening generalization. As this happens, the area around the winning units becomes tuned toward the two input stimuli, that is, many units' weight vectors are moved in the direction of the stimuli.

In essence, the mechanism outlined above results in an expansion of the area of the competitive layer that is devoted to the stimuli being discriminated. Interestingly, recent electrophysiological work from Merzenich and colleagues shows that perceptual learning is well-correlated with cortical representational changes in somatosensory and auditory cortex in that an increase in the amount of cortical space devoted to the discriminanda occurs with discrimination training (Jenkins, Merzenich, Ochs, Allard, & Guic-Robles, 1990; Merzenich, Recanzone, Jenkins, & Grajski, 1990). The above mechanism provides a way in which similar stimuli can eventually be discriminated. If the input stimuli are very different, however, separate competitive units will win from the start, so little perceptual learning will be seen.

A



B

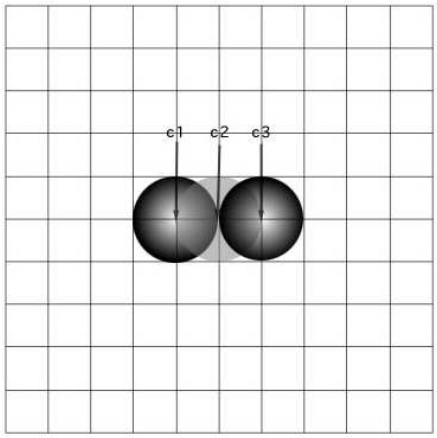


Figure 3-2: (A) Initial activation of the competitive layer after presentation of two similar stimuli, S_1 and S_2. The winning unit for S_1 is c_1 and the winning unit for S_2 is c_2. There is considerable overlap in the areas of activation triggered by each winner. Note that unit c_3 is in the neighborhood of c_2 but not c_1 and so is affected only by c_2. (B) After continued exposure to the stimuli c_3 becomes the new winner. This is because c_2 is in the neighborhood of c_1 and therefore it's weights are shifted toward c_1 when S_1 is presented. Since c_3 is not affected by the position of S_1 and its weights eventually become a more accurate representation of S_2.

Associative Processor

Each competitive unit has an association with the representation of an outcome (O) such as reinforcer or response. The strength of this association V_k is updated, using the delta rule (Rescorla & Wagner, 1972), for each competitive unit k proportionally to its activation.

$$\Delta V_k = \alpha_k \cdot \beta \cdot (\lambda - O) \tag{3-7}$$

In the above equation α_k is a variable learning rate associated with the competitive unit k, β is a static learning rate associated with the reinforcer, λ refers to the value of the reinforcer, and O is the outcome (in this case, the conditioned response of the animal to the stimulus). O consists of the sum of the associative strengths V_k of all of the competitive units scaled by their activation a_k.

$$O = \sum_k V_k \cdot a_k \tag{3-8}$$

The size of the increase or decrease in the association is dependent on the amount of attention α_k devoted to the stimulus configuration k. As mentioned, the rule for changing α is not directly addressed by the current model. Instead, the model incorporates a time-smoothed version of the well known associative mechanism, the Pearce-Hall rule (Pearce & Hall, 1980; Pearce & Kaye, 1985), to update α. As the outcome of the presentation of a stimulus becomes well predicted by the network (i.e., (λ - O) is minimized), α_k decreases in proportion to a_k.

$$\Delta\alpha_k = (1-\gamma) \cdot a_k \cdot \left[(\lambda - O) - \alpha_k (t-1)\right] \tag{3-9}$$

Interaction between the Nonassociative and the Associative Processor

Like associative learning, the degree to which tuning and separation occur on the competitive layer is affected by how well the outcome of the presentation of a stimulus is predicted. Thus, when the outcome is poorly predicted (e.g., at the beginning of discrimination learning or when the stimuli are difficult to discriminate) the degree of perceptual learning is high. As the outcome becomes better predicted, the learning rate on the competitive layer (δ) decreases according to the following rule:

$$\Delta\delta_k = (1-\gamma) \cdot a_k \cdot \left[(\lambda - O) - \delta_k (t-1)\right] \tag{3-10}$$

Note that the associative learning rate α and the competitive/perceptual learning rate δ are controlled by the same mechanism.

Behavior

Most of the experiments that will be presented in this document are simulations of simultaneous discrimination tasks. In this type of task, two stimuli are presented simultaneously and the animal must indicate its choice by pecking a key or pressing a touchscreen. On a given trial of a simultaneous discrimination simulation, stimulus A is input to the network and the activation of the input and competitive layers are computed. Then stimulus B is input to the network, and the activations again are computed. The initial output (O) produced by the model consists of the conditioned response (CR) to each of the presented stimuli. However, in order to reproduce as closely as possible the animal experiments, on each simulation trial two stimuli, A and B, are presented and the CRs to the stimuli are mapped to the probability of response to each stimulus (P(A) and P(B)) using the Boltzmann function:

$$P(A) = \frac{\exp(CR(A))}{\exp(CR(A)) + \exp(CR(B))} \tag{3-11}$$

For example, when the CR to each of the stimuli is near 0 at the beginning of training, P(A) = 0.50 and P(B) = 0.50. In the middle of training, on the other hand, if CR(A) = 0.60 and CR(B) = 0.30, P(A) = 0.77 and P(B) = 0.23. The actual response, either A or B, on a given trial is then chosen stochastically as a function of P(A) and P(B). To choose a response, a random number between 0 and 1 is generated. If P(A) is greater than the random number, then action A is chosen. If P(A) is less than the random number, then action B is chosen. Thus, if A is the correct stimulus, as P(A) increases, so will the likelihood that P(A) is greater than the random number and the increased CR(A) will be manifested in the network's choice behavior.

Summary

The present model concretizes the idea, originally presented by Channell and Hall (1981), that two separable processes—a reduction in associability and an increase in discriminability—occur during stimulus exposure. The associability of a stimulus, as manifested in its learning rate (α) is dependent on the amount of attention being paid to it. In the model, this form of attention is controlled according to the Pearce-Hall (1980) rule: attention to a stimulus decreases if its consequences are well-predicted and attention to the stimulus increases if they are not. In contrast with associability, the discriminability of a stimulus tends to increase with exposure. In the current model, this is due to a nonassociative, competitive learning mechanism that classifies stimuli according to their statistical properties. Over time, as the system acquires more input samples, representations of stimuli become more reflective of the actual stimulus parameters and thus become more precise. The particular conditions of a given experiment will influence the degree to which latent inhibition versus perceptual learning affect

General Methods

Identical network parameters were used for all simulations presented in this chapter. The network consisted of ten by ten grids of perceptual and competitive units. The specific parameters were as follows: $\alpha(0) = 0.05$, $\beta = 0.005$, $\delta(0) = 0.05$, $\gamma = 0.95$, $\lambda = 1.0$ (reinforced trials) or 0.0 (non-reinforced trials), $\sigma^2 = 3.0$.

The initial output produced by the network consists of the CR to the presented stimulus. For the concurrent discrimination simulations, the CR was translated to an actual stimulus choice as described in the Behavior section of this chapter. In this case, data are plotted as the percentage of trials in each block of 10 in which the network responded to the correct stimulus. Experiment 1 and Experiment 7, on the other hand, consist of conditioning to a single stimulus. In these simulations, the

behavioral response was considered to be a monotonic function of the CR, the value of which is plotted directly.

Error bars represent the standard error of the mean over 15 simulations, each of which can be said to represent a "subject" since the initial weights on the competitive unit links included a significant degree of randomness.

Experiment 1: Generalization

Guttman and Kalish's (1956) result that pigeons demonstrate a gradient of responding as a function of how similar a test stimulus is to the original training stimulus provides a clear demonstration of generalization. In this experiment, the authors reinforced pigeons for pecking at a key illuminated by an orange light. After training, the animals were tested with a variety of colors projected on the key, and the rate of responding to each color was recorded. They found that subjects responded most to the original stimulus, and the degree of response to other test stimuli declined in proportion to the difference, as measured in nanometers, between the test stimulus and the trained stimulus.

The current model accounts for generalization as a result of the topographical perceptual learning mechanism. When a stimulus is repeatedly presented, the weights of the winning unit are moved closer to the input pattern. The weights of neighbors are updated proportionally to their distance from the winner. As a result, a gradient of responding as a function of similarity to the trained stimulus should be seen when the network is trained on the Guttman and Kalish (1956) paradigm.

Methods

In this simulation, 15 networks were reinforced after presentation of a stimulus **S** = <5,5> for 100 trials. Their response to 10 stimuli of varying distance from the trained stimulus was then measured as an indication of generalization. On testing

trials, weights were not changed, which means that each network could be given all 10 tests without interference between them.

Results and Discussion

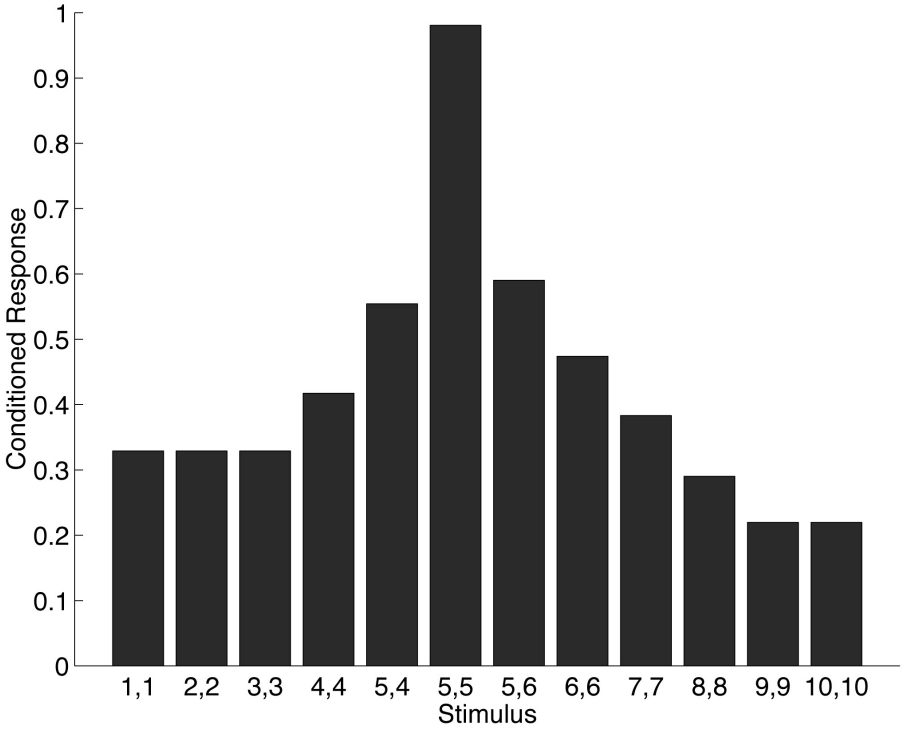


Figure 3-3: The generalization gradient produced by the network when it is trained on the Guttman and Kalish (1956) paradigm. In this simulation, 15 networks were reinforced for 50 trials after presentation of a stimulus S=<5,5>. Their response to 10 stimuli of varying distance from the trained stimulus was then measured as an indication of generalization. Comparable to Guttman and Kalish (1956), the distribution of the relative levels of responding is approximately Gaussian, with response level declining with distance from the trained stimulus.

Figure 3-3 shows the generalization gradient produced by the current model when it is trained on the Guttman and Kalish (1956) paradigm. Comparable to Guttman and Kalish (1956), the distribution of the relative levels of responding is approximately Gaussian, with the response level declining with distance from the trained stimulus. The steepness of the Gaussian is dependent on σ as well as the number of units on the grid.

This type of generalization is a direct result of the structure of the network. Since the neighborhood of the winning unit is Gaussian, one would expect to see a Gaussian generalization function. Therefore, whereas this result may not be surprising when one considers the structure of the network, it is important to verify that this effect is indeed produced by the model.

Experiment 2: Peak Shift

Köhler (1918), an early Gestalt psychologist, trained chickens to discriminate between cards of differing shades of gray. If the cards had been labeled in order from lightest to darkest, discrimination training would have been between shades 2(S+) and 3(S-). This phase was then followed by transfer tests to pairs of cards, some of which were of other reflectances (i.e., 1 vs. 2 and 3 vs. 4). What Köhler found was that during the transfer tests, the chickens tended to prefer 1 over 2. If they were trained with 3 as S+ and 2 as S-, however, they tended to prefer 4 to 3. Köhler suggested that this provided evidence that the animals were not learning associations between absolute reflectance and reinforcement. Instead, they were learning about a relationship between the S+ and the S- (i.e., one was brighter than the other).

In 1937, Spence presented an alternative to the relational point of view. He suggested that Köhler's result was due to an interaction between the different response tendencies acquired by the training stimuli. Thus, in discrimination training the S+ acquires excitatory tendencies whereas the S- acquires inhibitory tendencies. Response to subsequent test stimuli should consist of the algebraic sum of the tendencies. Spence showed that, assuming Gaussian generalization functions for the response tendencies, the maximum of their resultant occurs not at the S+, but at a point beyond it in the opposite direction of the S-. This prediction was confirmed experimentally by Hanson (1959) and subsequently termed a "peak shift". It has since been replicated numerous times (see Purtle, 1973 for a review).

The manner in which the current bottom-up model operates should produce an interaction similar to that proposed by Spence (1937). If two similar stimuli are presented to the network, the competitive layer neighborhoods of the winners for each stimulus will overlap. If the stimuli are differentially reinforced, it is probable that the winning competitive unit associated with the S+ will actually develop a smaller associative weight—due to S- generalized inhibition—than units in its neighborhood that are further from the S-. If this occurs, then on a subsequent test of conditioned response to the winner as well as the further units a peak shift should be evident.

Methods

In this simulation, 15 networks were trained on a discrimination between two inputs with overlapping neighborhoods of activation (S+ =<5,5> and S- = <5,6>). Then the response of each network to a range of stimuli was tested as a measure of generalization (<5,1>, <5,2>, <5,3>, <5,4>, <5,5>, <5,6>, <5,7>, <5,8>, <5,9>, <5,10>). On testing trials, weights were not changed, which means that each network could be given all 10 tests without interference between them.

Results and Discussion

Figure 3-4 shows the generalization gradient produced by the current model after discrimination training on two similar stimuli with overlapping areas of activation in the competitive layer. As the conditioned response was highest not for the S+, but for two stimuli in the neighborhood of the S+ in the opposite direction of the S- (<5,4> and <5,3>) the simulations reveal that the model produces a peak shift similar to that seen in the animal learning literature.

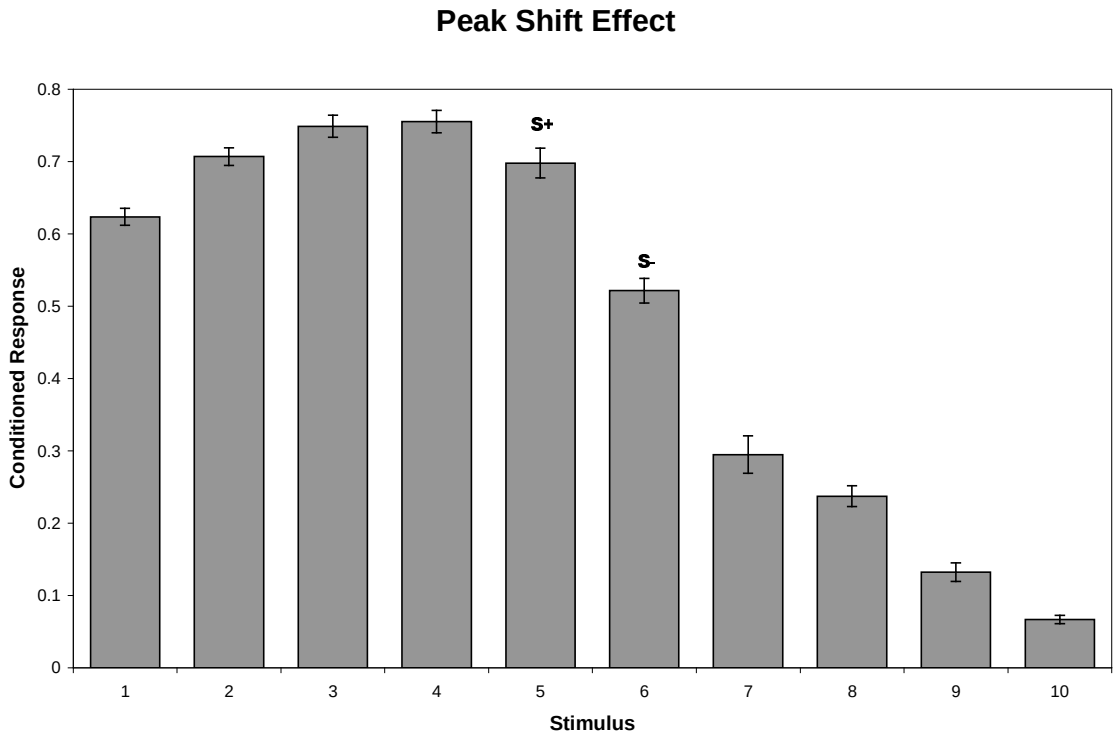


Figure 3-4: The peak shift effect.

As predicted above, the model produces peak shift as a result of overlap between the Gaussian areas of activation associated with the discriminanda. Because the winning units are situated within each other's area of activation, each of their associative strengths are both increased and decreased. A unit located within only the S+'s neighborhood, therefore, can develop greater associative strength than the winning unit due to the fact that it is only affected by the positive reinforcement of the S+ and not the negative reinforcement of the S-.

Differentiation moves the winning units further apart on the competitive layer, thereby reducing overlap of their activation neighborhoods. If the units are moved so far apart that there is little neighborhood overlap, there will be no interaction between response tendencies thus no peak shift will occur. However, since the competitive unit learning rate δ is decreased as trial outcome becomes well-predicted (see Equation 3-10), the winning units did not completely separate by the time this

discrimination .was learned to a reasonable criterion, and thus the peak shift was observed.

Experiment 3: Similarity

The argument that Pearce (1994) makes against elemental theories of discrimination learning is that they can incorrectly predict that making two patterns more similar will cause animals to learn the discrimination more quickly. Consider an experiment in which a common element, C, is added to two stimuli, A and AB, to produce AC and ABC. This manipulation presumably makes the two stimuli more similar due to their sharing of a common element. According to an elemental theory of learning based on the delta learning rule (e.g., Rescorla & Wagner, 1972), in a discrimination between AC+ and ABC-, on the first trial in which AC is presented, both A and C will gain associative strength. As a result, the overall associative strength of AC will be greater at the end of the first trial than if conditioning had been conducted with A alone. When ABC is presented, the delta rule predicts that the degree of negative associative strength acquired by B will be proportional to the overall associative strength of the compound. The presence of A and C will ensure that the magnitude of this associative strength is relatively high, and the increment in the negative associative strength of B will be greater than if it had been presented in an A+AB- discrimination. Thus, after the initial trials of the discrimination, the difference between the level of responding during AC and ABC will be greater than between A and AB. As a result, a discrimination between AC and BC should be learned more quickly than a discrimination between A and B. This prediction is not compatible with empirical results (e.g., Pearce & Redhead, 1993).

In this experiment I investigated whether making two stimuli more similar would cause the current model to learn the discrimination more slowly, as the data show, or more quickly, as an elemental theory would predict.

Methods

In this simulation, 15 networks were reinforced after presentation of a stimulus A=<2,2> and were not reinforced after presentation of a relatively different stimulus B=<8,8> (the Euclidean distance between A and B was 8.5). Next, 15 re-initialized networks were trained on a discrimination in which the stimuli were relatively similar (C=<5,5> and D=<5,6>; Euclidean distance =1.0).

Results and Discussion

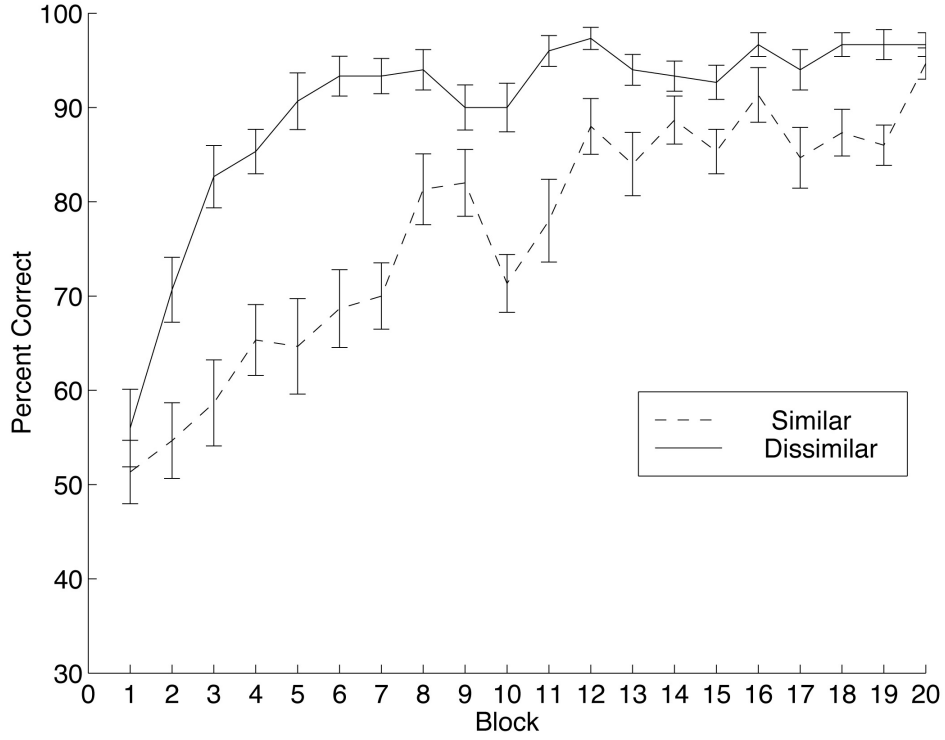


Figure 3-5: Discrimination ratios for two discriminations, one between relatively similar stimuli and one between relatively dissimilar stimuli. In this simulation, 15 networks were reinforced after presentation of a stimulus A=<2,2> and were not reinforced after presentation of a relatively different stimulus B=<8,8> (the Euclidean distance between A and B was 8.5). Next, 15 re-initialized networks were trained on a discrimination in which the stimuli were relatively similar (C=<5,5> and D=<5,6>; Euclidean distance =1.0).

The current model captures the fact that making stimuli more similar will decrease the rate at which they are discriminated. Figure 3-5 shows learning curves

for both the easy and the more difficult discrimination, averaged over 15 simulation runs. It is clear that the discrimination in which the stimuli were different was learned much more quickly than the discrimination in which the stimuli were more similar. In the current model, this occurs because different stimuli activate winning units that are further apart on the competitive layer than would be winners that similar stimuli activate. As a result, for different stimuli there is less overlap between active competitive units than there would be for similar stimuli. This leads to more consistent reinforcement and more rapid learning of the discrimination.

Experiment 4: Preexposure facilitates a difficult discrimination

The original Gibson-Walk experiment, along with others such as Oswalt (1972) and Chamizo and Mackintosh (1989) showed that relatively difficult discriminations are facilitated by preexposure. The current model's performance on a difficult discrimination, with and without preexposure, is tested in this experiment.

Methods

In this simulation, the stimuli presented were similar, with <*x,y*> coordinates of <5,5> for stimulus A and <5,6> for stimulus B. Two groups of 15 networks each (Group Exposed and Group Control) were trained with different procedures. During the first phase (preexposure), Group Exposed was given 100 interleaved presentations of each of A and B and Group Control was not run. During the second phase both groups were trained for 20 blocks of 10 trials on a discrimination in which A was reinforced and B was not.

Results and Discussion

The results of this simulation (see Figure 3-6) confirm that the model predicts facilitation of the subsequent discrimination for the similar stimuli that were used.

Initially, the winning competitive units coding A and B were located close to each other. This resulted in a good deal of overlap in the competitive layer activation functions for each stimulus. Thus when the two stimuli were repeatedly presented, a competition ensued in which each of the competitive units, when its corresponding input stimulus was delivered, moved it's weight vector as well as the other winner's weight vector toward the value of the stimulus. Because of this process, competitive units toward the far end of the neighborhood relative to the competing stimulus, which were less affected by the competing stimulus, gradually started to become winners. Over time, the two winning units became more and more separated on the configural grid, thus lessening their effect on each other and facilitating discrimination. In addition, each winning configural unit moved its neighbors closer, resulting in a reduced generalization gradient. Thus, for this discrimination the preexposure phase conferred an advantage because it reduced competitive layer overlap (corresponding to generalization) between patterns before the actual discrimination training phase began.

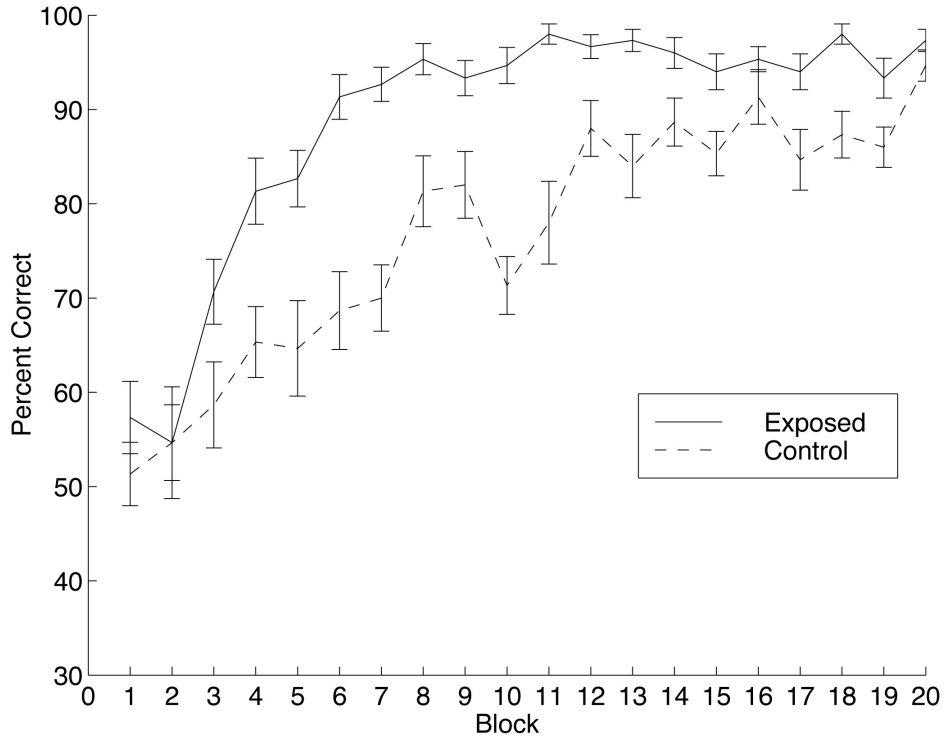


Figure 3-6: Results of a discrimination in which the stimuli presented were similar (A=<2,2> and B=<5,6>).During an initial phase (not shown here), Group Exposed was given 100 interleaved presentations of each of A and B and Group Control was not run. During a subsequent phase depicted in this figure, both groups were trained for 100 trials on a discrimination in which A was reinforced and B was not. Acquisition of the discrimination was faster for Group Exposed than for controls.

Experiment 5: Preexposure impairs an easy discrimination

Chamizo and Mackintosh (1989) also showed that in certain situations, specifically when a discrimination is not difficult, preexposure will retard a subsequent discrimination. The current model's performance on an easy discrimination, with and without preexposure, is tested in this experiment.

Methods

The procedure for this experiment was identical to that used in the previous simulation, with the exception that the stimuli used in this case were very different: the <*x,y*>coordinates of A were <2,2> and of B were <8,8>. Again, two groups of

15 networks each were run. Group Exposed were preexposed to the stimuli and Group Control were not.

The results of this simulation (see Figure 3-7) show that the model produces impairment of the subsequent discrimination for the relatively different stimuli that were used in this experiment. Initially, exposed animals were not impaired, but the learning curve for the control animals is steeper. This is because control subjects were subject to a lesser degree of latent inhibition and thus had a higher learning rate than the exposed group during discrimination training.

Results and Discussion

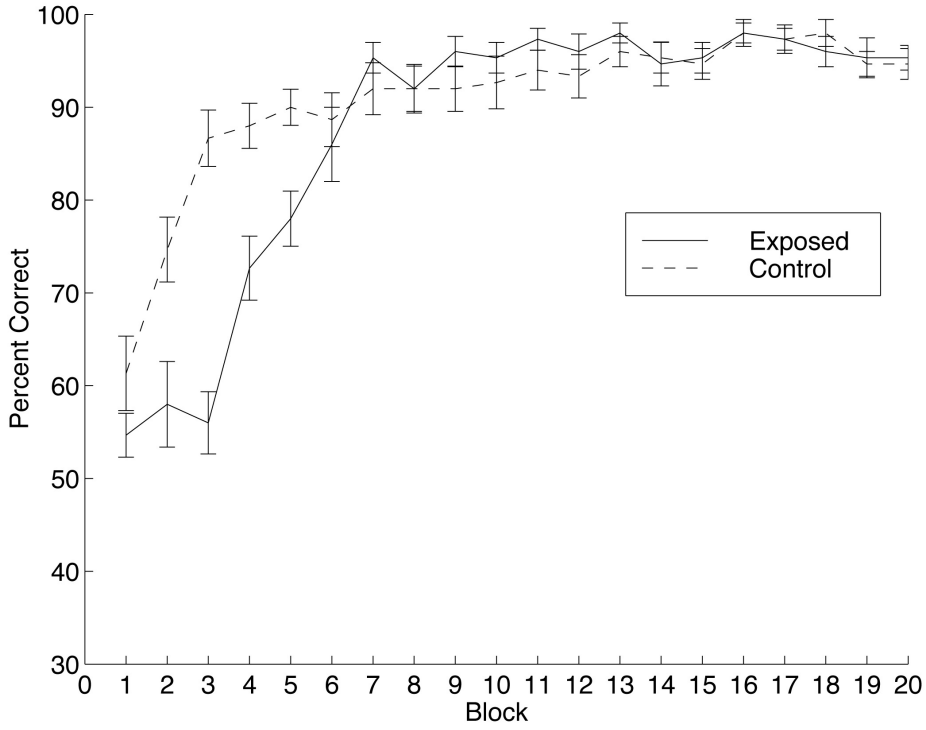


Figure 3-7:Results of a discrimination in which the stimuli presented were different (A=<2,2> and B=<8,8>). During an initial phase (not shown here), Group Exposed was given 100 interleaved presentations of each of A and B and Group Control was not run. During a subsequent phase depicted in this figure, both groups were trained for 100 trials on a discrimination in which A was reinforced and B was not. Acquisition of the discrimination was faster for controls than for Group Exposed.

Thus the current model predicts that impairment of subsequent discrimination will occur in the case of a relatively easy discrimination. Since the input patterns were far from each other in perceptual space, spatially distant competitive units were the winners from the start, thus there was little overlap in the competitive layer activation function for the two stimuli. In contrast to the situation in the previous simulation, in this case there was no advantage to preexposure. As a result, the only effect seen in the results here is that of latent inhibition.

Experiment 6: Separable latent inhibition and differentiation

In contrast to other models of perceptual learning (e.g., McLaren et al., 1989), in the current model the mechanisms of latent inhibition and differentiation are independent and separable. Thus it is possible to show each of their effects separately in order to demonstrate their operation and to show that the seemingly inconsistent effects of preexposure to stimuli are simply due to an interaction between the two processes.

Methods

In this simulation, 30 networks were altered such that the latent inhibition mechanism, but not the perceptual learning mechanism, was operational. Then 15 of the networks were preexposed to two similar stimuli (A=<5,5> and B=<5,6>) while the remaining 15 networks were not. Next, all 30 networks were trained on a discrimination between A=<5,5> and B=<5,6>. The networks were then reinitialized and 15 were exposed to two different stimuli (A=<2,2> and B=<8,8>) while the remaining 15 were not. Finally, all 30 networks were trained on a discrimination between A=<2,2> and B=<8,8>.

In the second phase of this experiment, 30 newly initialized networks were altered such that the perceptual learning mechanism, but not the latent inhibition

mechanism, was operational. Then 15 of the networks were preexposed to two stimuli (A=<5,5> and B=<5,6>) while the remaining 15 networks were not. Next, all 30 networks were trained on a discrimination between A=<5,5> and B=<5,6>. The networks were then reinitialized and 15 were exposed to two different stimuli (A=<2,2> and B=<8,8>) while the remaining 15 were not. Finally, all 30 networks were trained on a discrimination between A=<2,2> and B=<8,8>.

Results and Discussion

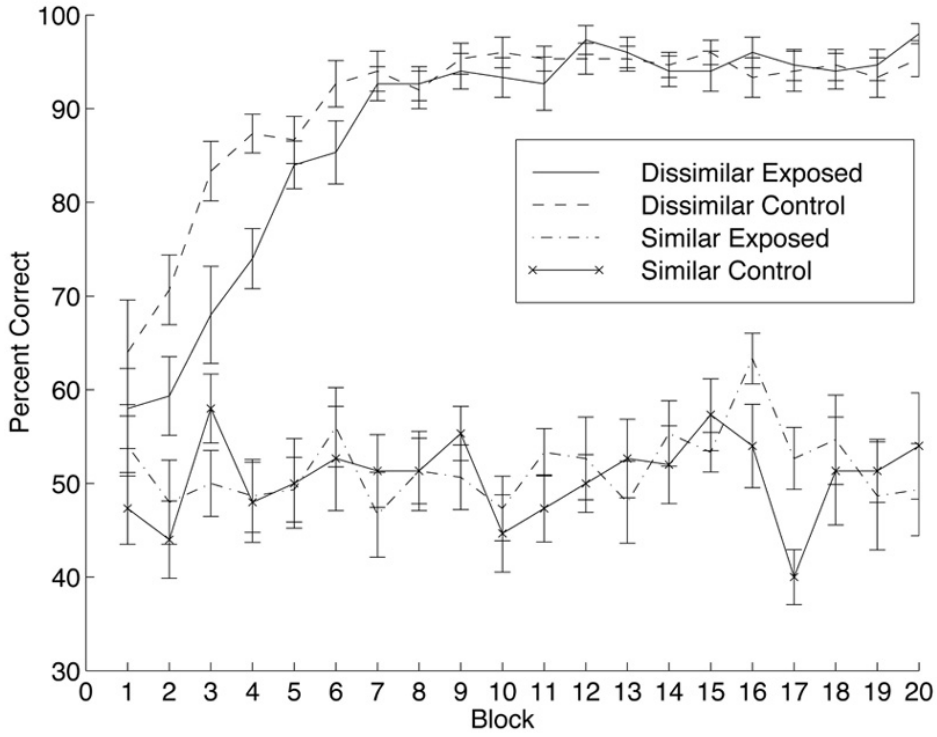


Figure 3-8: The effect of preexposure on an easy and a difficult discrimination when only the latent inhibition mechanism is in effect.

Figure 3-8 illustrates the effect of preexposure on the discriminations used in this experiment when only the latent inhibition mechanism is in effect. Neither group was able to learn the difficult discrimination. Since the network had been able to learn this discrimination in an earlier experiment (see Experiment 3), this was presumably due to the absence of the discrimination mechanism. On the easy

discrimination, the exposed group learned more slowly than the controls, thereby indicating a latent inhibition effect.

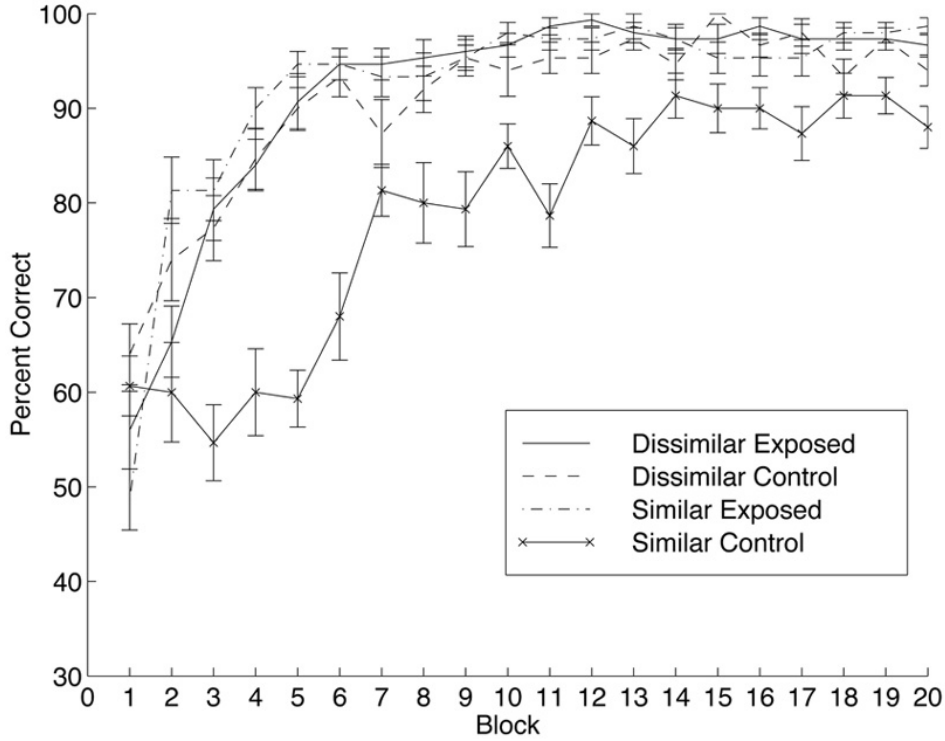


Figure 3-9: The effect of preexposure on an easy and a difficult discrimination when only the differentiation mechanism is in effect.

Figure 3-9 shows the effect of preexposure on the same discriminations when only the perceptual learning mechanism is active. Quite a different pattern of results is seen here. Both discriminations are learned by all of the groups, however the control group's acquisition of the difficult discrimination is much slower than the other three groups. There is no difference in acquisition between the other three groups. This suggests that preexposure to stimuli would help with any discrimination, easy or difficult, if latent inhibition weren't counteracting the positive effects.

Thus the current model consists of two completely separable and independent mechanisms for learning during exposure to a stimulus: Latent inhibition reduces the learning rate of preexposed stimuli while differentiation enhances their discriminability. The fact that these two processes operate simultaneously suggests that one should be able to pull the two mechanisms apart in a conditioning experiment. Honey and Hall (1989) demonstrated just such an effect.

Experiment 7: Simultaneous latent inhibition and perceptual learning

Honey and Hall (1989) demonstrated simultaneous effects of latent inhibition and perceptual learning in the same animals. Rats were preexposed to separate presentations of each of two distinctly flavored solutions (A and B). Control rats were given water. Next, all subjects received a series of trials in which A was followed by LiCl, thereby conditioning an aversion to A. Acquisition of the aversion developed more slowly for preexposed animals, presumably due to latent inhibition. Discrimination between A and B was then assessed with a test in which B was presented and its generalized tendency to evoke a CR was measured. Animals in the control group tended to generalize to and thus reject B, whereas those in the A/B preexposure group showed little generalization from A to B. These results suggest that preexposure slowed conditioning to A while facilitating discrimination between A and B. Thus, this experiment provides support for the idea that there are two opposing forces operating during preexposure.

In this experiment, a simulation of Honey and Hall (1989) was used to investigate whether the simultaneous effects of perceptual learning and latent inhibition could be seen within the current model.

Methods

Two groups of 15 networks each were run in this simulation of Honey and Hall (1989). Group Exposed were preexposed to 100 presentations of A <5,5> and B <5,6>. Then both groups were conditioned to A to a criterion of CR=0.95. Finally, both groups were presented with B for one trial in extinction.

Results and Discussion

Figure 3-10 shows trials to criterion for acquisition of a CR to A for the preexposed and control groups. As in Honey and Hall (1989), Group Exposed were impaired on acquisition of the response with respect to controls. This reflects the influence of the latent inhibition mechanism on the exposed group.

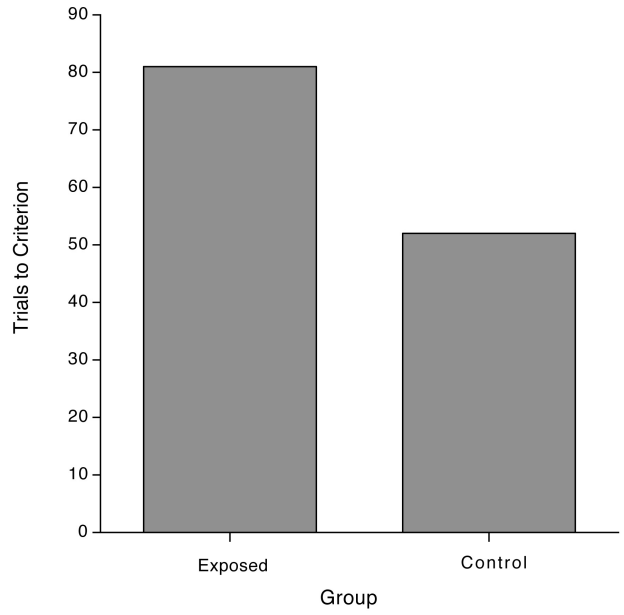


Figure 3-10: Trials to criterion for a simulation in which Group Exposed were preexposed to 100 presentations of A<5,5> and B<5,6> whereas controls were not preexposed. Then both groups were conditioned to A to a criterion of CR=0.95. Group Exposed were impaired on acquisition of the response with respect to controls.

Performance for the two groups on the extinction test is shown in Figure 3-11. The extent of the CR on this test indicates the degree of generalization from B to the reinforced stimulus A. The group means were 0.73 for group Control and 0.42 for group Exposed, indicating that the preexposed group showed less generalization than the control group. Thus even though the associative learning rate of Group Exposed was lowered as a result of preexposure, stimulus discriminability was increased.

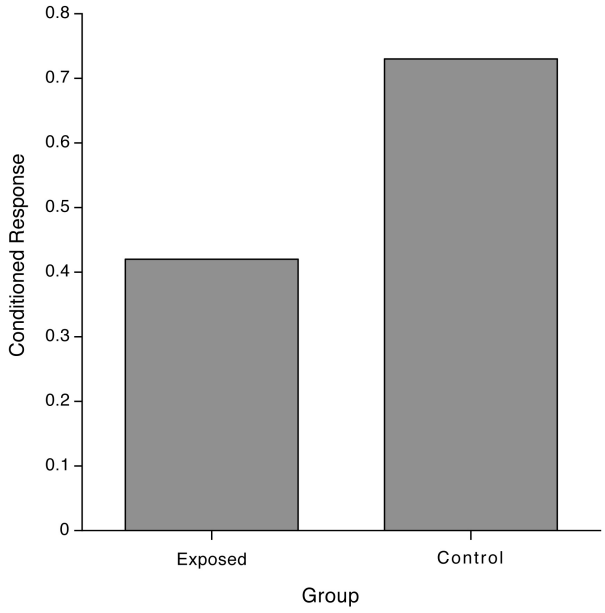


Figure 3-11: Results from the generalization test in the simulation of Honey & Hall (1989). Group Exposed was exposed to 100 presentations of A<5,5> and B<5,6> whereas controls were not preexposed. Then both groups received 100 training trials in which A was reinforced. Finally, both groups were presented with B for one trial in extinction. The preexposed group showed less generalization than the control group.

The above simulation shows that the model produces simultaneous latent inhibition and perceptual learning, just as in Honey and Hall (1989). Preexposed subjects acquired the CR more slowly than controls, but they also generalized less to

the reinforced stimulus. This is due to the fact that the model comprises separate mechanisms for associability and discriminability which operate independently.

General Discussion

The current chapter describes a connectionist model of perceptual learning that accounts for data showing that preexposure can lead either to facilitation or retardation of subsequent learning. This is achieved through the incorporation of a concrete, nonassociative differentiation mechanism within an associative learning framework. In the present model the behavioral effects of preexposure are directly dependent on the difficulty of the discrimination. If two input stimuli are very similar then they will tend to be captured by nearby units on the competitive layer. With more input samples, however, the perceptual learning mechanism causes better separated units eventually to win out over the original winners. This results in the stimulus representations on the competitive layer being pushed apart; as a result preexposure will provide an advantage on a subsequent discrimination learning task. If the input stimuli are very different, however, well-spaced competitive units win from the start, so advance differentiation confers no advantage. In both cases, latent inhibition of the preexposed stimuli also occurs during preexposure. Because differentiation is unnecessary in the latter case, however, latent inhibition effects tend to be much more apparent.

One factor that sets the current model apart from other models of perceptual learning is that it is compatible with a configural approach to associative learning without using biologically implausible backpropagation of error (cf.McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). The present chapter has focused on the basic perceptual and discrimination learning mechanism by restricting inputs to unidimensional stimuli that can be represented in a single input layer. However, it is fairly straightforward to see how the model could be extended in order to accommodate multi-dimensional or compound stimuli in a configural manner. For example, imagine that a compound stimulus AB is presented to the network. As in

the basic network discussed in this paper, stimuli would consist of vectors of values; however in this case they would be quadruples (e.g., <*x1,y1,x2,y2*>) rather than pairs (<*x1,y1*>), with an <*x,y*> value within the quadruple for each of A and B. In order to accommodate this increased number of stimuli, two separate input layers ($\mathbf{I_A}$ and $\mathbf{I_B}$) would be utilized. Each input layer would represent one stimulus, either A or B. Thus if a compound stimulus were input to the network, each of the input layers would be activated depending on the value of the stimulus to which it corresponded. The pattern of activation on the input layers would be generated in the same manner as described for the basic perceptual and discrimination learning mechanism. In order for the network to be configural, the competitive portion of the network would be fully connected with both $\mathbf{I_A}$ and $\mathbf{I_B}$ and therefore would be able to represent the conjunction of the stimuli or features represented in these layers. The competitive layer, with its configural representation of the input stimuli, would be fully connected with the outcome representation. In sum, the full configural version of this model would be similar to that proposed by Pearce (1994) with the exception that instead of being made up of linked single units it would be made up of linked grids of competitive units. Thus, the configural version of the current network would have the same functionality as Pearce's (1994) network, with the added benefit that it would be able to deal with changing stimulus representations as in perceptual learning.

A second unique aspect of the current model is that it suggests that perceptual learning is based on a singular, nonassociative mechanism that focuses on the separation of the representations of stimuli. Thus, the current model suggests that associative factors may not be strictly necessary for perceptual learning (although they may in certain circumstances influence the processing of stimulus representations during preexposure, see Bennett, Wills, Wells, & Mackintosh, 1994; Espinet, Iraola, Bennett, & Mackintosh, 1995; Mackintosh, Kaye, & Bennett, 1991; Trobalon, Chamizo, & Mackintosh, 1992; Trobalon, Sansa, Chamizo, & Mackintosh, 1991). This contrasts strongly with elemental models in which associative mechanisms are the foundation of perceptual learning. The model of McLaren et al.

(1989), for example, suggests that three associative processes contribute to perceptual learning. First, stimulus sampling theory (Neimark & Estes, 1967) is extended by adding the assumption that in addition to associations between elements of different events, associations also form between elements comprising a single event, such as the CS itself. Variance in the input is derived from the fact that only a subset of the elements is activated on each trial. Over time, associations will form between the most frequently sampled elements in a process called "unitization", thus creating an elaborated stimulus representation that reflects the central tendency of the sampled elements. Second, the theory suggests that common elements are subject to twice as much exposure as unique elements during preexposure thus they are subject to twice as much latent inhibition. As a consequence, preexposure reduces the subsequent rate of associative learning for common elements relative to unique elements, thereby facilitating discrimination. Third, since the stimuli will be perceived separately, inhibitory associations will form between unique elements.

Summary

The present model reconciles associative and nonassociative accounts of preexposure effects in discrimination learning by incorporating aspects of both approaches in a single model. It focuses on a nonassociative perceptual learning mechanism that is based on the idea that stimulus representations can be tuned and separated during simple exposure. Unlike other perceptual learning approaches, the model is compatible with a configural approach to associative learning. In addition, since it is instantiated in a computational model, it provides concrete mechanisms behind what have to this point been presented only as abstract theoretical concepts (e.g., Gibson & Gibson, 1955).

CHAPTER 4 DIMENSIONAL ATTENTION

Introduction

Many researchers have argued that discrimination learning involves, in addition to the low-level effects discussed in Chapter 3, changes in higher-level attentional processes. For example, in an early experiment Lawrence (1949) investigated whether transfer between tasks might occur on the basis of increased stimulus discriminability. In this experiment, rats in the critical experimental condition were trained on a simultaneous black-white discrimination. Control subjects were trained on a discrimination in which the maze arms were an identical shade of gray, but the textures of the floors differed. The basis of the transfer test that followed was to make the associations learned in the first stage irrelevant: in some trials both ends of the maze were black, signaling that a response to the right (say) was rewarded; in other trials both ends were white, signaling that a response to the left was rewarded. Thus, having a tendency to approach black rather than white would not aid acquisition of this new task, but an increase in the distinctiveness of black and white might. Indeed, animals in the experimental condition did learn the successive discrimination more readily than controls, and this result has been confirmed by subsequent experiments (Jaynes, 1950). Although several interpretations of these data are possible [2], Lawrence's interpretation was based on the relatively high-level concept of attention: Lawrence suggested that increases in discriminability are due not to an increase in the ease of discrimination between two stimuli (as one would expect with a low-level perceptual learning mechanism), but instead to an increase in the ease with which the cues enter into further associations due to changes in attention. After training on a black/white discrimination, he suggests, the animal is

[2] For example, see the acquired distinctiveness argument presented at the beginning of Chapter 5.

not any better at discriminating black from white. Instead, the dimension of brightness comes to command the animal's attention better than previously.

An equally plausible interpretation of these data, however, is that the effect is due to altered low-level stimulus representations rather than to high-level changes in attention. In other words, the facilitation that Lawrence found might be due not to greater attention to the brightness dimension, but to greater separation between the discriminanda through perceptual learning. This explanation, if it turned out to be feasible, would be interesting for at least two reasons: (1) Differentiation, as modeled by the current network, can account for a wide variety of perceptual learning data that are not accounted for by an attentional explanation and (2) Differentiation is a low-level mechanism and thus may be more parsimonious than high-level attentional mechanisms. As such, in this chapter I investigate the extent to which data usually attributed to dimensional attention can be explained simply as a result of differentiation.

In the following section, some classic studies of attentional transfer after discrimination learning are described. Next, theoretical accounts of these phenomena are discussed. Finally, I present a series of simulations suggesting that many effects heretofore attributed to dimensional attention may in fact be more parsimoniously explained by the current model of differentiation.

Evidence for Attentional Processing in Discrimination Learning

In most investigations of dimensional attention experimenters attempt to separate attentional transfer effects from associative, or direct, transfer effects. One paradigm often used for separating these processes is reversal learning. In a reversal learning experiment, animals are first trained on a discrimination task and then in a second "reversal" stage the reinforcement contingencies are reversed. Usually, as one might expect, learning in the second stage is impaired as a result of associative transfer. However, Reid (1953) found that rats given an extra 150 trials on a simultaneous black- white discrimination learned the reversal more quickly than

subjects trained only to a moderate criterion. This suggests that some process other than associative transfer is in effect. One possible account similar to that provided by Lawrence to explain his (1949) experiment: during a black-white discrimination there is a strengthening of the tendency to attend to brightness plus a reduction in tendency to attend to other dimensions along which the cues differ. This shift in dimensional attention outweighs the direct transfer effects in the overtrained animals thereby leading to a facilitation in the reversal test.

A second way in which researchers have addressed the issue of direct versus attentional transfer is by leaving the response requirement the same in the training and transfer test but changing the stimuli such that the effect of direct transfer is reduced, whereas the effect of attentional transfer is preserved. One example of this is the intra-dimensional (ID) shift. In an ID shift, the specific stimuli are changed when the subject is shifted from one stage to the next but the dimension along which the stimuli differ is kept the same. With appropriate stimuli there should be no reason for associative transfer to give a subject a preference for one stimulus over the other. But if the first stage of training increases attention to the relevant dimension, then there should be positive transfer from the first task to the second. The control condition consists of an extra-dimensional (ED) shift in which subjects are moved to a test discrimination in which the stimuli differ along a completely different dimension than in the training stage. An example of this type of experiment was done by Mackintosh and Little (1969). Subjects were trained on a discrimination in which stimuli consisted of lines that could differ in color or orientation. Group 1 was trained on a discrimination in which red horizontal or vertical lines were reinforced while green horizontal or vertical lines were not. Group 2 was trained on a discrimination in which red and green vertical lines were reinforced while red and green horizontal lines were not. Then both groups were subjected to a transfer test in which blue left-diagonal lines and blue right-diagonal lines were reinforced and yellow left-diagonal and right-diagonal lines were not. Thus, Group 1 experienced an ID shift on the transfer test whereas Group 2 experienced an ED shift. Group 1 learned the test problem more quickly than Group 2, suggesting that more transfer

occurs across an ID shift. This effect can also be explained by an increase in attention to the relevant dimension, and Mackintosh (1974) suggests that this is "perhaps the best evidence that transfer between discrimination problems may be partly based on increases in attention."

A different approach to intradimensional transfer results from work on transfer along a continuum. Lawrence (1952) trained rats using stimuli consisting of shades of gray ranging from light (A) to dark (G) with five intermediate shades. Rats trained from the outset on a difficult discrimination (D-E) learned slowly. Others that were given initial training trials on the easiest task (A-G) performed very well when given the ID shift to the D-E task and their terminal level was superior to that of the first group. An analysis based on Lawrence (1949), would suggest that initial training strengthened the relevant dimension and this led to positive transfer. Mackintosh and Little (1970) looked at the possibility that the effect might instead be due to generalization of associative strength. Subjects were pigeons that were going to have to learn a very difficult wavelength discrimination. They were given easy pretraining with the values of the stimuli reversed. Generalization of associative strength would put this group at a disadvantage. They found this to be the case at the beginning of the test phase: subjects given reversed pretraining were inferior to those trained on the difficult discrimination from the beginning. As training continued, however, the pretrained group overcame its disadvantage and overtook the group trained only on the hard task.

Theoretical Interpretations

In this section of the chapter I describe two major theories of dimensional attention. I have chosen these models in particular because they are well-formulated relative to other theories in the literature (e.g., Lawrence, 1963). The first, analyzer theory (Mackintosh, 1975; Sutherland & Mackintosh, 1971), is the predominant model of dimensional attention effects in the animal learning literature (Hall, 1991). The second, ALCOVE (Kruschke, 1992), is a connectionist model of categorization

and discrimination learning based largely on Nosofsky's generalized context model (1986) from the human cognitive psychology literature.

Analyzer Theory

Sutherland and Mackintosh (1971) presented "analyzer" theory, a formalization of Lawrence's idea that discrimination training might produce a change in attention as reflected in stimulus associability. Like the current model, analyzer theory suggests that discrimination learning is a two-stage process in which stimulus representations pass through a stage of perceptual learning before associations are constructed between it and other events. However, unlike the current model, analyzer theory suggests that the perceptual learning consists of changes in attention to the dimension to which the stimulus belongs. Thus, as Lawrence (1963) originally suggested, in a black-white discrimination, there is a strengthening of the tendency to attend to brightness (called the brightness analyzer) plus a reduction in the tendency to attend to other dimensions along which the cues differ. In order for an analyzer to gain strength, the outputs of the analyzer must be consistently related to certain consequences: i.e., in the above example black reliably predicts food and white predicts its absence. When the outputs of the analyzer fail to make correct predictions about the outcome, its strength decreases. In addition, Sutherland and Mackintosh suggest that the capacity of the dimensional attention system is limited such that an increase in strength of attention to one analyzer causes a decrease in others. This change in analyzer strength is assumed to be fundamentally different from the associative changes that occur between representations of stimuli and reinforcement[3].

[3] Mackintosh (1975) later proposed a revised version of analyzer theory. In this revision, it was proposed that changes in analyzer strength are specific to stimuli instead of dimensions. Second, it was suggested that analyzer strength is increased when a stimulus successfully predicts its consequences and is decreased when it is unsuccessful. The third change was that adjustments in analyzer strength happen independently; the assumption of a limited capacity system was dropped. However, in this chapter discussion centers around the original version of analyzer theory.

ALCOVE

Kruschke (1992) presents a connectionist model of category learning (ALCOVE) that accounts for several experimental results in the human categorization literature. The model incorporates an explicit representation of continuous stimulus dimensions with an attentional parameter associated with each dimension, and determines which dimensions are most important to a given task and how strongly to associate exemplars with a given category. The mechanism underlying the model consists of an exemplar-based representation modeled on the Generalized Context Model (Nosofsky, 1986) that is paired with backpropagation learning (Rumelhart et al., 1986).

A brief summary of the model is as follows. Input nodes represent psychological dimensions, and the activation of a node indicates the value of the particular stimulus on that dimension. The input nodes are gated by a dimensional attention strength that reflects the relevance of the dimension for the categorization task at hand. Each hidden node consists of a point in a multidimensional psychological space and represents a particular exemplar. The activation of a given hidden node is proportional to the similarity between the stimulus and the exemplar, and this activation falls off exponentially with the distance between the hidden node and the input stimulus. The dimensional attention strengths function as multipliers on the dimension when computing the distance between a stimulus and the hidden node exemplar, so the attention strengths functionally stretch and shrink dimensions of input space such that stimuli in different categories are better separated and stimuli within categories are closer together. A linear rule adjusts associative strengths between exemplars and categories and category activations are mapped onto response probabilities using a Boltzmann function choice rule. Thus for each presentation of a training exemplar, activation propagates to the category nodes. The teacher values are then presented, compared with actual category node activation, and association strengths are adjusted in proportion to the degree of error. The error is then backpropagated to adjust the dimensional attention weights.

The model can account for some classic data in the human categorization literature. Kruschke (1992) demonstrates the application of ALCOVE to several studies. The study most relevant to this dissertation is that of Shepard, Hovland, and Jenkins(1961), in which learning to attend to relevant dimensions was investigated. In the experiment, stimuli varied on three binary dimensions: shape (square vs. triangle), size (large vs. small), and color (filled vs. open). The resulting eight exemplars were evenly divided into two categories. It turned out that there were six structurally distinct types of category assignments, with Type I requiring information only about dimension 1, Type II requiring attention to dimensions 1 and 2, and Types III, IV, V, and VI requiring information about all three dimensions in order correctly to categorize the stimuli (although the dimensions are not all equally informative in each type). As a result, if it takes more cognitive effort to learn about more dimensions, then Type I category assignments should be the easiest to learn, followed by the remainder of the types. The empirical result supported this notion, and Shepard and colleagues (1961) concluded that extant learning theories could not account for the data. In particular, the authors measured interstimulus similarities and then computed the difficulty of category types by considering the similarities of all pairs of exemplars from different categories. In so doing, they found that a generalization hypothesis—the idea that categorization structures that assign similar stimuli to the same category and dissimilar stimuli to different categories should be relatively easy to learn—did not predict the correct order of category type difficulty. As a result, they proposed the idea of selective attention: devotion of attention to only the relevant dimensions leads to an increase in discriminability between stimuli that differ on those dimensions. Because ALCOVE is based on the learning of dimensional attention weights, it is well-formulated to describe data of this type, and as such has no difficulty accounting for these results.

ALCOVE provides a clear demonstration of the manner in which dimensional attention may shift according to task demands. It can also account for other categorization phenomena such as the fact that similar examples from the same category should influence each other's learning. However, like analyzer theory, the

model is based on the assumption of the existence of explicit representations of stimulus dimensions that can be attentionally weighted. In addition, the model assumes that attention to the entire dimension is shifted depending on the relevance of stimuli. As a result, it cannot address issues regarding local changes in attention for a given dimension.

Accounting for Dimensional Attention Effects without Attention

As outlined above, current models of "dimensional attention" effects such as Sutherland and Mackintosh (1971) and Kruschke (1992) assume the existence of explicit dimensional representations that are weighted by attentional values. While these models have provided a good characterization of dimensional attention phenomena, it seems unlikely that the singular dimensional representations upon which they rely exist in the brain (e.g., a representation of "color"). In addition, while it is certainly possible that this type of dimensional attention mechanism operates simultaneously with a competitive perceptual learning mechanism, the principle of parsimony suggests that it would be useful to see to what extent the differentiation model alone is sufficient to account for the dimensional attention data.

Toward this end, I would like to propose that "dimensional attention" effects are at least partially a consequence of differentiation mechanisms operating on networks of neurons that are organized in a specific manner. This proposal is based on two fundamental assumptions derived from the properties of real neurons:

1. Different neurons respond differentially to different stimulus dimensions. A vast amount of evidence from neuronal recording studies exists to suggest that this is the case (e.g., Desimone, 1996; Sakai, Naya, & Miyashita, 1994; Tanaka, 1997). Certain neurons respond best to edges, others to specific colors, others to certain shapes, etc.

2. Neurons are topographically organized according to the dimension to which they respond best. Again, evidence from neuronal recording experiments supports this assumption. Tanaka (1997), for example, has found that neurons responding primarily to color are located in area TE in the macaque temporal lobe and that neurons responding to more complex conjunctions of features may be located in perirhinal cortex (PRh). This type of topographic organization of neuronal response properties seems to be a general property of visual cortex organization (Kandel, Schwartz, & Jessell, 1991).

Based on the above assumptions I would like to suggest that dimensional attention cannot be captured by a single parameter attached to a representation of dimension. Instead, the effect of dimensional attention might *emerge* from perceptual learning mechanisms operating on representations of individual stimuli. I do not intend to suggest that no representation of dimension exists. Instead, I would like to claim that these representations exist in a distributed manner across the response properties of neuronal populations. As such, dimensional attention effects are due not to a single parameter, but instead are the result of a combination of perceptual learning effects and the generalization of neuronal response between stimuli.

In order to test this idea, I ran a series of simulations of experiments that have been used as evidence for the existence of dimensional attention mechanisms. The model used for these simulations consists of mechanisms for the type of perceptual learning known as differentiation and is described fully in Chapter 3. The model contains no explicit representation of dimensional attention.

General Methods

Identical network parameters were used for all simulations presented in this chapter, except where noted. The network consisted of ten by ten grids of perceptual and competitive units. The specific parameters were as follows: $\alpha(0)=1$, $\beta=0.005$, $\delta(0)=$, $\gamma=0.95$, $\lambda=1$ (reinforced trials) or 0 (nonreinforced trials), $\sigma^2(0)=2.0$.

The initial output produced by the network consists of the CR to the presented stimulus. Each of the following experiments involves simultaneous discriminations, so the CR was translated to a stimulus choice as described in the Behavior section in Chapter 3. All data are plotted as the percentage of trials in each block in which the network responded to the correct stimulus. Block sizes differ depending on the difficulty of the discrimination, and are indicated in the individual experimental methods sections.

Experiment 1: Overtraining Reversal Effect (ORE)

To recap, the ORE reflects the finding that animals that are overtrained on a discrimination can learn a subsequent reversal more quickly than less well-trained controls. This effect has been explained by theories of dimensional attention as follows: animals that are overtrained pay more attention to the stimulus dimension relevant to solving the discrimination. This attentional effect outweighs any negative transfer caused by the reversal of reinforcement contingencies. In this experiment I tested whether the ORE could be explained by considering the idea that the phenomenon could be a result of simple differentiation rather than being caused by an attentional shift.

Methods

In this simulation, two groups of 25 networks each, group Control and group Overtrained, were initialized. Both groups were trained to a criterion of 100% correct for three trials in a row on a difficult one-pair simultaneous discrimination. Group Overtrained was then trained on the same discrimination problem for an additional 50 trials. Next, both groups were trained on a reversal of the original problem in which the original S+ was no longer reinforced and the original S- was now reinforced.

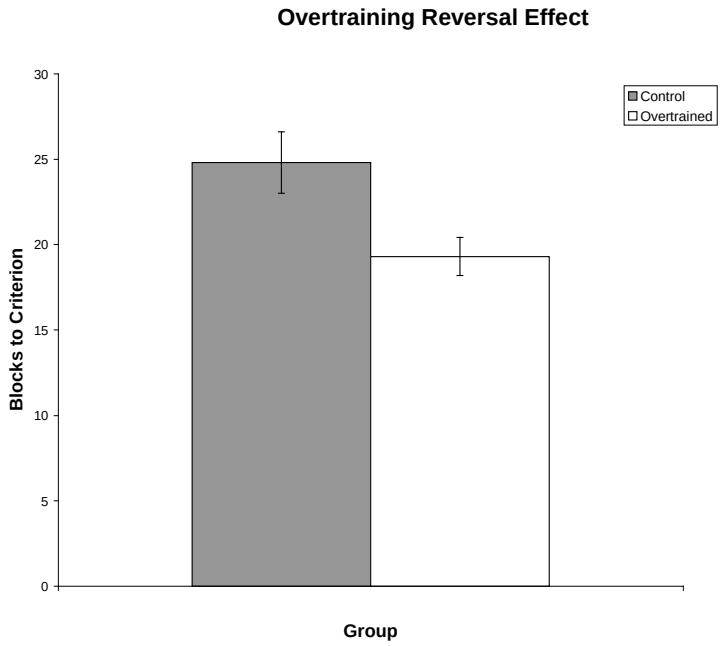


Figure 4-1: Performance of Group Control and Group Overtrained on a reversal of the original discrimination. Group Control were trained to a criterion of 100% correct on three trials in a row on the original discrimination, and were then tested on the same problem with the reinforcement contingencies reversed. Group Overtrained were trained to the same criterion as Group Control, and then were given an extra 50 discrimination trials before being transferred to the reversal problem. Group Overtrained learned the reversal more quickly than Group Control.

Figure 4-1 shows performance of the two groups on the reversal test. Group Overtrained learned the reversal more quickly than Group Control, suggesting that a notion of dimensional attention is not necessary to account for the ORE. The current model's mechanism of differentiation is sufficient.

The model produces the ORE as a side effect of the differentiation mechanism. During initial discrimination training, in addition to acquiring associative strength, stimulus representations are also separated on the competitive layer. When the discrimination is learned to a moderate criterion, such as that used for the control group in this experiment, the stimulus representations become fairly well-separated and the network is able to learn the discrimination reversal fairly quickly. However, when the initial discrimination is overtrained, the stimulus representations become

very well-separated, and this actually facilitates learning of the reversal because the advantage gained as a result of the increased pattern separation outweighs the disadvantage caused by the increased associative transfer.

Experiment 2: Transfer of Learning Across a Continuum

The topographical organization of the competitive layer along with the fact that neighbors of winning units are updated proportionally to their distance from the winner leads the model to the following prediction: exposure to a pair of similar stimuli will facilitate not only discrimination between the exposed pair, but also discrimination between stimuli that are even more similar. This is because two stimuli that are not easily discriminable will initially be represented as nearby winning units in the competitive layer. As the stimuli are presented to the network, the winning units will be pulled apart and each of the units between them will become tuned toward the stimulus to which it is closest. This tuning of intermediate units suggests that preexposure to two stimuli will facilitate discrimination of other stimuli whose representations fall between them on the competitive layer.

Some support for the above prediction derives from studies of discrimination transfer along a continuum, a method used to investigate intradimensional transfer. Lawrence (1952) found that rats trained from the outset on a discrimination between two stimuli lying close together on a brightness continuum learned this difficult discrimination rather slowly. A second group of rats that were initially trained on an easy task in which the stimuli differed greatly, and then were transferred to the difficult discrimination, learned the difficult discrimination much more quickly than the first group of rats. Lawrence's finding has proven to be robust, with a similar effect having been seen in species ranging from humans (Gonzalez & Ross, 1958; Hogg & Evans, 1975; May & MacPherson, 1971) to honeybees (Walker, Lee, & Bitterman, 1990).

As discussed earlier in this chapter, the mechanism surmised to underlie the above experimental results is usually one of dimensional attention. For example,

(Lawrence, 1952) suggested that animals pretrained on an easier version of the discrimination were more quickly able to determine the relevant stimulus dimensions for solving the discrimination: it is likely easier to determine that brightness is important in a black vs. white discrimination as compared to a discrimination between two shades of gray. However, if the current model can reproduce this effect, this would suggest that reference to the concept of dimensional attention is not necessary to explain transfer along a continuum. Thus in this experiment I tested, using Lawrence's (1952) methodology, whether the model could produce transfer along a continuum.

Methods

In this simulation, two groups of 25 networks each, group Transfer and group Control, were initialized. Group Control was trained for 10 blocks of 50 trials each on a very difficult discrimination between two stimulus pairs, A=<5,5> vs. B=<5,5.25>. Group Transfer was trained on a simpler version of the same problem (X=<5,4.5> vs. Y=<5,5.75>) for two blocks of 50 trials and was then shifted to the same problem as group Control (A=<5,5> vs. B=<5,5.25>) for 8 blocks of 50 trials each. Thus each group experienced the same total number of training trials.

Results and Discussion

Figure 4-2 shows learning curves for both group Control and group Transfer, averaged over 25 training runs. Clearly, group Transfer was able to learn the discrimination more quickly than group Control. This confirms that the model predicts an advantage in learning if the networks are initially trained on a simpler version of a discrimination between two very similar patterns.

The reason that the model produces this behavior is as follows. Group Control was trained on the difficult discrimination from the start. As a result, the patterns were initially categorized as being the same since they activated the same winning competitive unit. Due to the competitive updating, the winning units eventually separated around Block 5, at which point the discrimination started being performed

above chance. Around Block 8, the winners separated further, leading to an increase in the rate of learning of the discrimination as a result of reduced overlap between associative links. Group Transfer, on the other hand, was initially trained on a discrimination in which the patterns were further separated than those in the target discrimination. The networks were able immediately to separate on the competitive layer the patterns of the initial discrimination, and thus began learning the discrimination quickly. When they were shifted to the more difficult discrimination, separate units on the competitive layer won for the two stimuli because of changes in the competitive layer that had occurred as a result of the initial discrimination. Thus the Transfer group learned the difficult discrimination much more quickly than the Control group.

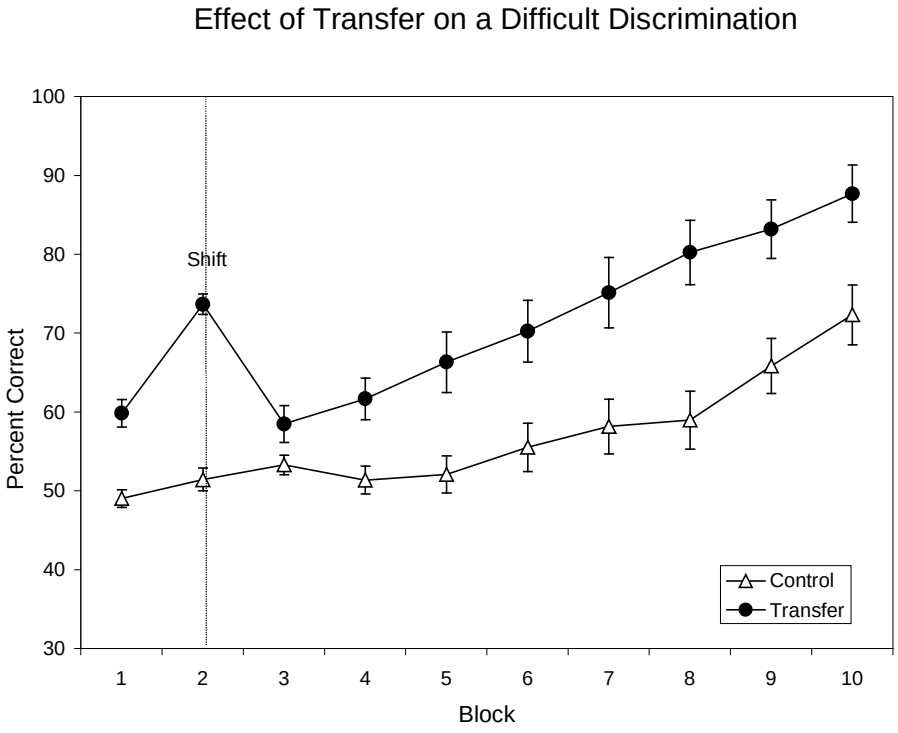


Figure 4-2: Learning curves for a discrimination between very similar patterns (A=<5,5> vs. B=<5,5.25>). Group Transfer were trained on a simpler version of the discrimination for one block, whereas group Control were trained on the difficult discrimination from the start. Although the groups were trained for the same total number of trials, group Transfer acquired the discrimination more quickly.

The simple fact that this model can produce the effect of transfer along a continuum suggests that the notion of dimensional attention is not necessary to explain this phenomenon. Earlier researchers (Lawrence, 1952) argued that the transfer along a continuum phenomenon indicates that theories based on stimulus generalization gradients are insufficient as they assume that transfer is greatest between identical discriminations. That is, for a fixed amount of practice on a training discrimination, the transfer effect to a test discrimination will be greatest when the training and test discriminations are identical. Lawrence further suggested that generalization gradient-based theories must postulate unique types of generalization gradients and make unlikely assumptions about the additive properties of habit strength. Earlier discrimination theories based on generalization gradients, however, did not take into account perceptual learning and as such had no mechanism for separation of patterns. The current model is based on stimulus generalization gradients but, since it was designed to simulate perceptual learning, can also account for the effect of transfer along a continuum via a single mechanism without having to assume additional concepts of dimensional attention.

Experiment 3: Further Analysis of Transfer Along a Continuum

One possibility is that transfer along a continuum is due merely to associative transfer (see Logan, 1966). In other words, it could be the case that the improvement in the transfer group in Lawrence (1952) was not due to an improvement in perceptual distinction between the stimuli, but to the fact that during the initial, simpler discrimination the animal built up an association between the S+ and reinforcement, and this transferred to the S+ in the more difficult discrimination. Fairly strong evidence against such a view, however, was subsequently reported by Mackintosh and Little (Mackintosh & Little, 1970). In this study, pigeons were trained on an easy wavelength discrimination prior to a more difficult one, but the values of the stimuli were reversed between training phases. That is, if the longer of

the two wavelengths was rewarded during the initial training phase, then in the subsequent phase the shorter wavelength was rewarded. The authors demonstrated that subjects given reversed pretraining were initially worse on the difficult discrimination task, but they eventually overtook the group that was only trained on the hard task. Mackintosh and Little suggested an attentional explanation: during pretraining, the appropriate analyzer (Sutherland & Mackintosh, 1971) becomes established, and this effect is significant enough to outweigh any negative transfer from the generalization of associative strength.

The current model suggests an alternative interpretation of the above result that does not require postulation of an additional attentional mechanism. Specifically, the pretrained group will benefit from the tuning and separation of the competitive units, but at the same time will suffer from generalization of associative strength between the new and the pretrained winners. These two effects will counteract each other, and the behavior observed will depend on the relative strength of each of them. Thus as the winning units become more separated, generalization of associative strength between them will decrease, and performance will improve. This property, combined with the advantage of extra separation between the new winners gained during pretraining, should be sufficient for the current model to reproduce the pattern of data observed by Mackintosh and Little (1970) without having to make additional assumptions about dimensional attention.

Methods

In this simulation, two groups of 25 networks each, group Reversed Transfer and group Control, were initialized. Group Control was trained for 10 blocks of 50 trials each on a difficult discrimination with <5,5> as the S+ and <5,5.25> as the S-. Group Reversed Transfer was trained on an easier version of the same problem with S+ and S- reversed (<5.75,5> as the S+ and <5,4.5> as the S-) for two blocks of 50 trials. In other words, the stimulus in the easy discrimination that was most similar to the S+ in the difficult discrimination was not reinforced whereas the stimulus in the easy discrimination that was most similar to the S- in the difficult discrimination was

reinforced. The Reversed Transfer group was then shifted to the same problem as group Control (S+=<5,5> vs. S-=<5,5.25>) for 8 blocks of 50 trials each. Thus each group experienced the same total number of discrimination learning trials.

Results and Discussion

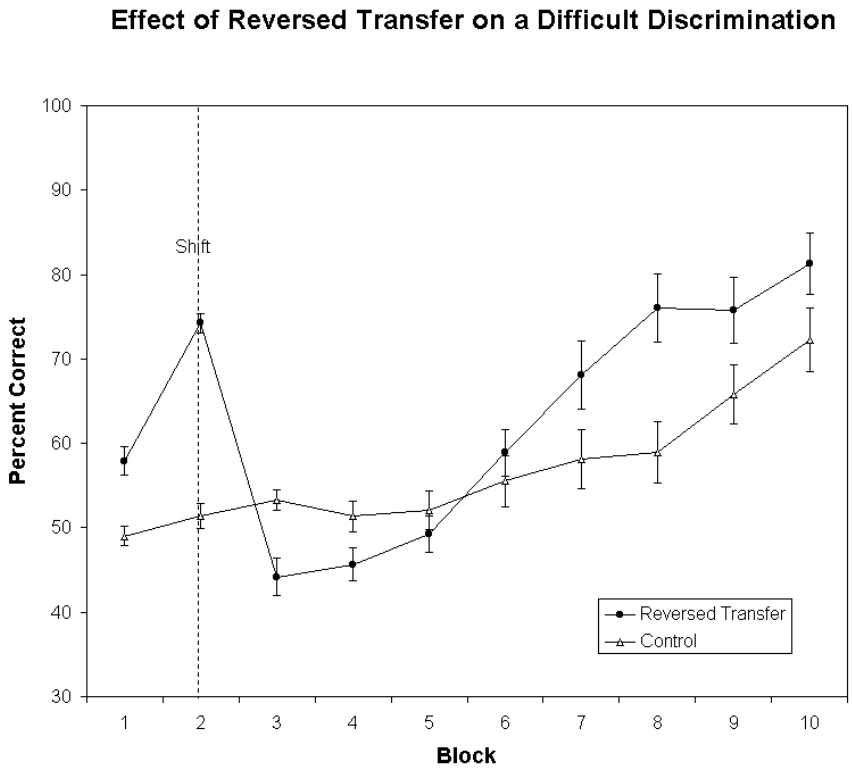


Figure 4-3: Learning curves for a discrimination between very similar patterns (A=<5,5> vs. B=<5,5.25>). Group Reversed Transfer was trained on a simpler version of the discrimination, with the S+ and S- reversed as compared to the final discrimination, for one block. Group Control was trained on the difficult discrimination for the entire session. Although the groups were trained for the same total number of trials, and associative transfer for group Reversed Transfer impeded learning of the final discrimination for several blocks, overall group Reversed Transfer acquired the discrimination more quickly.

Figure 4-3 shows learning curves for both group Control and group Reversed Transfer, averaged over 25 training runs. Group Reversed Transfer demonstrated a severe drop in performance when the final difficult discrimination was introduced at the beginning of Block 3. Because the S+ and S- were reversed in the initial discrimination relative to the S+ and S- in the final discrimination, associative

transfer worked against group Reversed Transfer, resulting in worse performance of the transfer group relative to controls during blocks 3 through 5. However, due to the changes made in the competitive layer as a result of training on the initial discrimination, networks in group Reversed Transfer were able to learn the new associations more quickly and so overtook controls and overall were able to learn the discrimination more quickly. This pattern of data, with the reversed transfer group showing an initial decrease in discrimination followed by superior learning performance as compared to controls, mirrors the results found by Mackintosh and Little (1970).

Like in the previous simulation of "straight" transfer along a continuum, Reversed Transfer networks were able immediately to separate on the competitive layer the patterns of the initial discrimination, and thus began learning the discrimination quickly. When they were shifted to the more difficult discrimination, separate units on the competitive layer won for the two stimuli as a result of changes in the competitive layer that had occurred as a result of the initial discrimination. Thus changes in the competitive layer facilitated discrimination of the stimuli. At the same time, however, the S+ had developed negative links with reinforcement whereas the S- had developed positive associative links. This negative associative transfer caused the Reversed Transfer networks initially to perform worse on the difficult discrimination than controls. However, the advantage gained through further separation on the competitive layer of the stimulus representations outweighed the negative influence of associative transfer, and the Reversed Transfer networks outperformed controls in the long run.

Mackintosh and Little (1970) suggested that their finding that reversed training on the easy discrimination was effective in producing transfer along a continuum implies that the beneficial effects of an easy problem are caused more by the establishment of strengthening of attention to the relevant dimension than by the establishment of differences in response strength to the test stimuli. While it is true that associative transfer is not a good explanation of the effect, the fact that the current model can reproduce the effect demonstrates that the notion of dimensional

attention is not necessary to explain these data. Instead, transfer along a continuum may be due to changes in the stimulus representations that occur as a result of initial exposure to the easy discrimination.

Experiment 4: Comparison of Similar Stimuli

Several recent studies have shown that the opportunity to compare stimuli during preexposure seems to affect the subsequent ease of discrimination. For example, when similar stimuli are preexposed in a mixed fashion (i.e., A then B then A then B) subsequent discrimination between A and B is facilitated more than when stimuli are exposed in a blocked fashion (i.e., AAAA then BBBB, Honey et al., 1994). Symonds and Hall (1997) report a similar advantage of mixed preexposure to two stimuli A and B over preexposure to B alone. For example, consider the experiment of Honey and colleagues (1994) in which chicks were preexposed to two relatively similar stimuli in either a mixed or a blocked fashion. In the case of mixed preexposure, chicks were given 50 presentations of stimulus A intermixed with 50 presentations of stimulus B. In the case of blocked preexposure, chicks were first presented with 50 trials of stimulus A which were followed by 50 trials of stimulus B. When both groups were subsequently trained on a discrimination between A and B, the mixed group outperformed the blocked group[4].

The above data are in line with Gibson's claim that the opportunity to compare two stimuli should facilitate discrimination as "Simultaneous comparison is no doubt the simplest for differentiation of two stimulus objects and the discovery of contrasts and feature differences must begin in this way."(Gibson, 1969, p. 145). Unfortunately, whereas Gibson's hypothesis appears to be supported by the data, Gibson has provided no concrete mechanism that might produce the effect. The current model accounts for stimulus preexposure effects by assuming that they are due to alterations of the stimulus representations on the competitive layer. This

[4] Interestingly, the opposite is true for relatively easy discriminations: Blocked preexposure facilitates subsequent discrimination more than does mixed preexposure. See Experiment 5 for discussion.

mechanism presents the possibility that different methods of preexposure, such as mixed and blocked, may cause different types of changes in the stimulus representations on the competitive layer, which may in turn affect the rate of subsequent discrimination learning. An attentional explanation for these data is more difficult to imagine. Thus, in this experiment I tested whether mixed and blocked preexposure would in fact differentially alter the competitive layer and lead to differences in subsequent discrimination learning, thereby suggesting a possible mechanism for the stimulus comparison effect.

Methods

In this simulation, two groups of 25 networks each, group Mixed and group Blocked, were initialized. Group Mixed was preexposed for four blocks of 10 trials with interleaved presentations of two stimuli, <5, 5> and <5, 5.25>. Group Blocked was preexposed to two blocks of 10 trials of stimulus <5, 5> followed by two blocks of 10 trials of stimulus <5, 5.25>. Thus, both groups received the same number of stimulus presentations during preexposure, but for the first group the two stimuli were interleaved whereas for the second group the stimuli were not. Both groups were then trained on a discrimination in which the S+ was <5, 5> and the S- was <5, 5.25>.

Results and Discussion

Figure 4-4 shows learning curves for Group Mixed and Group Blocked on a discrimination between two similar stimuli. Although the effect was small, overall group Mixed acquired the discrimination more quickly than group Blocked. This result is consistent with that shown by Honey et al. (1994; note that, like the simulation, the size of the effect in that experiment was small), and can be explained by the following behavior of the model.

In the current model, during the blocked trials in which A is presented, the initial winning competitive unit (cA_1) remains as the winner throughout. During this time weights in the competitive layer become tuned toward A and the competitive

learning rate of A decreases. As a result of this tuning, when B is presented, the new winning unit (cB_1) is further from cA_1 than it would have been without preexposure. During the B preexposure phase, because cB_1's competitive learning rate is relatively high, cB_1 and its neighbors become tuned toward B. However, since cA_1's competitive learning rate is low, it is not influenced by the presentation of B, and it remains tuned toward A. As a result, when mixed stimulus presentations then occur during discrimination training, presentation of stimulus A continues to produce the same winner, cA_1. Thus, blocked preexposure facilitates subsequent discrimination as a consequence of the initial tuning of weights toward A, which causes cB_1 to be further from cA_1 than it would have been without preexposure. In the case of mixed preexposure, however, the winners constantly compete and the representations of A and B are pushed apart on the competitive layer. Unlike the blocked preexposure case, there is no period during which one unit consistently wins until its learning rate decreases significantly, causing the stimulus representation to become static. Under the current parameters, this means that the mixed exposure condition will lead to further separation of the winning competitive units, and thus better discriminability, than blocked preexposure.

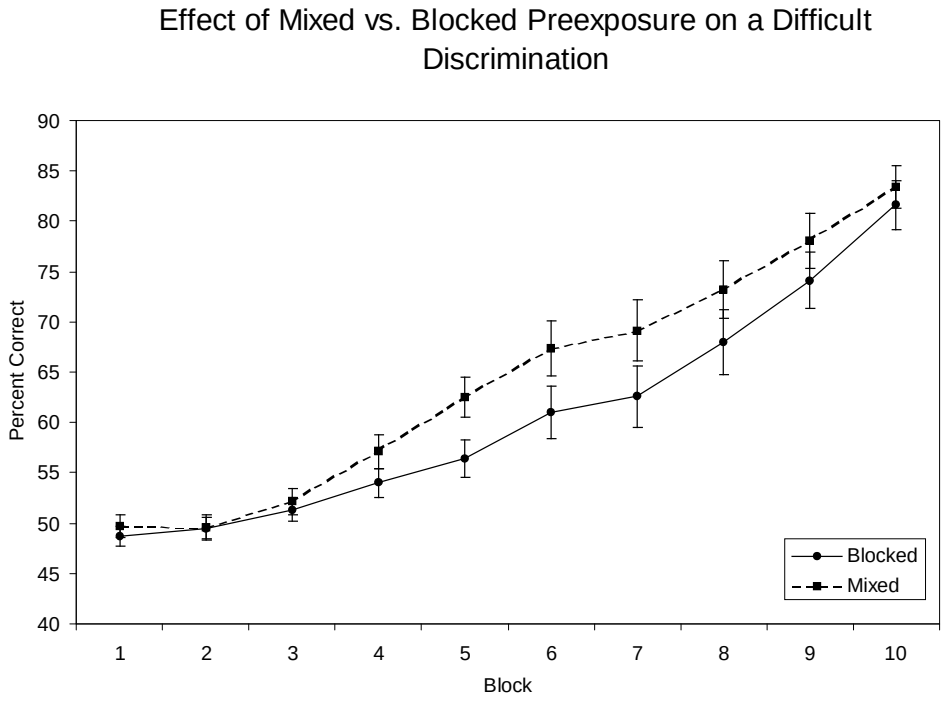


Figure 4-4: Learning curves for a discrimination between very similar patterns (A=<5, 5> vs. B=<5, 5.25>). Group Blocked were preexposed to the discriminanda in a blocked fashion (i.e., 20 presentations of stimulus A followed by 20 presentations of stimulus B). Group Mixed were preexposed to the discriminanda in a mixed fashion(i.e., 40 interleaved presentations of A and B). Although the effect was small, overall group Mixed acquired the discrimination more quickly than group Blocked.

The success of the current simulation suggests that interaction between changing stimulus representations on the competitive layer is a potential mechanism for production of the stimulus comparison effect. The relatively slow development of the stimulus representations, combined with the integration of the representations through interleaved presentation may provide a more balanced representation of the stimuli that allows them to be more easily discriminated. This method may also reduce the likelihood of interference between stimuli.

Further investigation of mixed and blocked preexposure is warranted, however, as the effect in animals seems to depend on the difficulty of the stimuli: when Honey et al. (1994) ran the same experiment using a more discriminable stimulus pair they found that subjects that had been preexposed in a blocked fashion were *facilitated* on

a subsequent discrimination relative to those that had had mixed exposure. Thus, the effect of mixed versus blocked preexposure on a simpler discrimination is explored and discussed in the following experiment.

Experiment 5: Comparison of Easily Discriminable Stimuli

A second paradox appears in the discrimination learning literature in the context of mixed versus blocked stimulus preexposure. Honey and colleagues, in the same 1994 paper discussed earlier, found that in certain cases blocked stimulus preexposure facilitates subsequent discrimination more than does mixed preexposure. Analysis of the discrimination problems suggests that the effect seems to depend on the difficulty of the discrimination: Difficult discriminations are facilitated most by mixed exposure whereas easier discriminations are facilitated most by blocked exposure.

Honey et al. explain their data by suggesting that two processes are at work during preexposure. First, it is well-known that temporal contiguity promotes association. Thus, associations between the preexposed stimuli are most likely to have developed during the mixed sessions, when A and B were relatively more contiguous. This provides a straightforward explanation of performance on the easy discrimination task. Second, at the same time a differentiation process operates to develop stimulus representations. This process separates and refines the stimulus representations during preexposure, thereby possibly facilitating performance on subsequent discriminations. Since the benefit of comparison is seen most when the discriminanda are similar (see Chapter 3), it only makes sense that differentiation outweighs association when the discrimination is difficult but not when it is easy. These two mechanisms provide a convincing explanation of the data.

As seen in the previous experiment, the current model can reproduce the effect of mixed versus blocked preexposure on a difficult discrimination. However,

according to the model, on an easy discrimination mixed and blocked preexposure should produce effects similar to each other: Because the competitive units should be well-separated from the start there should be no interaction between the units that could lead to a difference in learning between the two types of preexposure. In order to verify this prediction, I replicated Experiment 4 using a much more discriminable pair of inputs.

Methods

In this simulation, two groups of 10 networks each, group Mixed and group Blocked, were initialized. Group Mixed was preexposed for four blocks of 10 trials with interleaved presentations of two stimuli, <5, 4> and <5, 8>. Group Blocked was preexposed to two blocks of 10 trials of stimulus <5, 4> followed by two blocks of 10 trials of stimulus <5, 8>. Thus, both groups received the same number of stimulus presentations during preexposure, but for the first group the two stimuli were interleaved whereas for the second group the stimuli were not. Both groups were then trained on a discrimination in which the S+ was <5, 4> and the S- was <5, 8>.

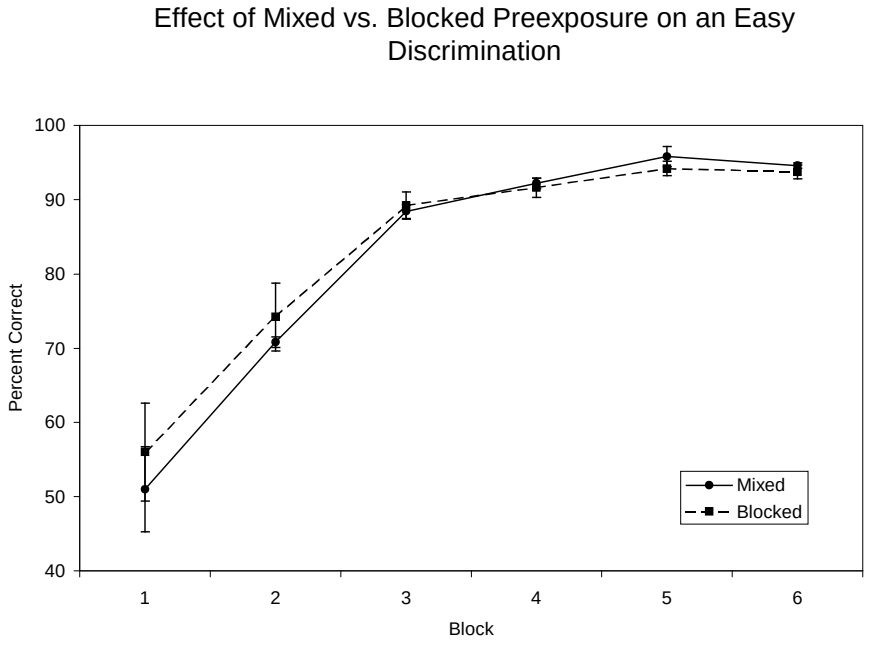


Figure 4-5: Learning curves for a discrimination between relatively different patterns (A=<5, 4> vs. B=<5, 5.8>). Group Blocked were preexposed to the discriminanda in a blocked fashion (i.e., 20 presentations of stimulus A followed by 20 presentations of stimulus B). Group Mixed were preexposed to the discriminanda in a mixed fashion(i.e., 40 interleaved presentations of A and B). There appears to be no difference between groups on this manipulation when the stimuli are easily discriminable.

As expected, there was no difference between group Mixed and group Blocked on learning the easy discrimination after the two types of preexposure (see Figure 4-5). For an easy discrimination it is simply much less likely that one would see a differential effect between mixed and blocked exposure because the initial winning competitive units are so far apart that there is not likely to be any interaction between them. Thus, whereas the current model provides a differentiation mechanism that can account for the facilitatory effect of mixed exposure on a difficult discrimination, it cannot account for the superior effect of blocked exposure on an easy discrimination.

Honey et al. attribute the inferior performance of the mixed group on the easy discrimination to interference between stimuli as presenting stimuli close together in

time results in their becoming associated with one another. The idea of associative links forming between contiguous stimuli has much support in the literature both generally (Domjan & Burkhard, 1986) and with respect to preexposure (Honey & Bateson, 1996). As the current model is trial-based, however, it cannot account for any temporally-dependent effects. This was clearly demonstrated in the current simulation, and as such points out an apparent shortcoming of the current model. In future work it might be very useful to develop a finer-grained version of this model which would allow for exploration of non trial-based effects. It seems that the added complexity that such an undertaking would entail would be worthwhile in the face of the additional phenomena that might be investigated and accounted for.

Experiment 6: Further Analysis of Stimulus Comparison

As shown in Experiment 5, the stimulus comparison data can be accounted for by the differentiation mechanism of the current model. The stimulus comparison data are equally consistent with at least one mechanism—mutual inhibition—postulated by McLaren and colleagues (1989). These authors suggest that the inhibition that can develop during mixed preexposure will tend to be different from that which develops during blocked preexposure. In the former case, the inhibition will be mutual in that the unique elements of A will inhibit the unique elements of B and vice versa. In the latter case, only the unique elements of the stimulus presented in the second session will inhibit the unique elements of that presented in the first. This difference in inhibition could lead to an advantage for subjects given mixed stimulus preexposure.

Thus in order to distinguish between the two proposed mechanisms for stimulus comparison one must devise an experiment to test mixed versus blocked preexposure in which the likelihood of mutual inhibition is minimized. If an advantage of mixed preexposure is demonstrated while the likelihood of mutual inhibition is low, then this would provide evidence against a mutual inhibition account of stimulus comparison, and suggest that an alternative theory may be

needed. A hint in this direction is provided by Saldanha and Bitterman (1951). In this experiment, two pairs of stimuli were used, one consisting of two differently shaded cards (A and B), the other consisting of two vertically striped cards (X and Y) in which the stripes were of different widths. One group of rats ("paired") were initially trained on two concurrent discriminations in which stimuli differing along the same dimension were compared (A+ vs. B- and X+ vs. Y-). A second group ("unpaired") were trained on two concurrent discriminations in which stimuli differing along different dimensions were compared (A+ vs. Y- and X+ vs. B-). Once each group attained criterion on their first set of discriminations, they were switched to the other set of discriminations. The authors found that the unpaired group was impaired, on both sets of discriminations, relative to the paired group.

These data have been taken to imply the use of relational cues—for example, "A is brighter than B"—during very difficult simultaneous discriminations. Honey et al. (1994) argue that these data are contrary to what a mutual inhibition account of stimulus comparison would predict: for the unpaired group, mutual inhibition between the unique elements, and associative strength between the common elements, should be greater than in the paired group. As a result, according to a mutual inhibition account of stimulus comparison, the unpaired group of animals should have performed better than the paired group. In this experiment, I tested whether the current model makes a false prediction as would a mutual inhibition-based model.

Methods

In this simulation, two groups of 25 networks each, group Paired and group Unpaired, were initialized. Group Paired was trained to a criterion of 80% correct on two pairwise concurrent discriminations in which the stimuli varied on the same dimension[5](S+=<5,0>, S-=<6,0>;S+=<0,6>, S-=<0,5). This was followed by

[5] A dimension is assumed to correspond to one axis of the two-dimensional stimulus representation in the input layer and the competitive layer.

training on two pairwise concurrent discriminations in which the stimuli varied along different dimensions (S+=<5,0>, S-=<0,5>;S+=<0,6>, S-=<6,0). Group Unpaired was trained on the same discriminations in reversed order. That is, Group Unpaired was trained first on the unpaired discrimination (S+=<5,0>, S-=<0,5>;S+=<0,6>, S-=<6,0) and second on the paired discrimination (S+=<5,0>, S-=<6,0>;S+=<0,6>, S-=<0,5).

Results and Discussion

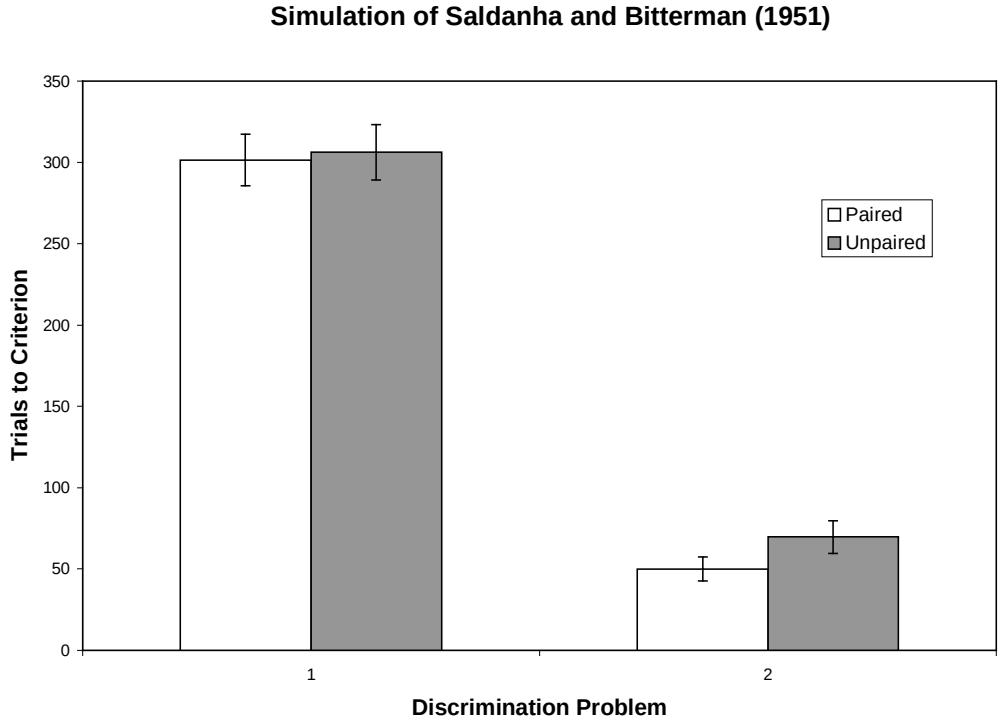


Figure 4-6: Trials to criterion on two discrimination problems. Group Paired was first trained to a criterion of 80% correct on Discrimination 1, in which the stimuli varied along the same dimension, and was then trained on Discrimination 2, in which the stimuli varied along different dimensions. Group Unpaired was trained first on Discrimination 2 and then on Discrimination 1. There was no significant difference between Group Paired and Group Unpaired on performance of either problem.

Figure 4-6 shows performance as measured by trials to criterion for Group Paired and Group Unpaired on two discrimination problems. Group Paired was first trained on Discrimination 1, in which the stimuli varied along the same dimension,

and was then trained on Discrimination 2, in which the stimuli varied along different dimensions. Group Unpaired was trained first on Discrimination 2 and then on Discrimination 1. There was no difference between Group Paired and Group Unpaired on performance of either problem.

Unlike the McLaren et al. model, the current model does not make the incorrect prediction of an advantage for the unpaired group. At the same time, however, the current model does not predict the advantage for the paired group that was found by Saldanha and Bitterman (1951). Thus the current model, at least in its present form, is insufficient to account for this particular experiment. Two reasons for this seem likely. The first problem is with basic stimulus representation. In this experiment I used a two-dimensional representation with the x axis representing one dimension and the y axis representing a second dimension. This particular representation may not be able to capture the different stimulus dimensions in a reasonable manner, thus the interactions between paired and unpaired may not make sense. In order to investigate this possibility further, I re-ran the current experiment using a more developed stimulus representation that is described in Chapter 6. However, the results were very similar to those presented in Figure 4-6, and thus are not presented here. Even with a more complex stimulus representation, there was no difference between the paired and unpaired groups, suggesting that the impoverished stimulus representation is not the reason that the current model did not replicate Saldanha and Bitterman's (1951) data.

A second, more important, problem with this simulation is that there was essentially no difference between the paired and the unpaired stimulus presentations in terms of how the stimulus representation weights in the network were altered. In Saldanha and Bitterrman's (1951) experiment, the same stimuli were presented in both conditions, the only difference between groups being the timing of their presentation. Although the organization of the experiment and this simulation are the same, because of the lack of temporal information in the model's framework it does not really matter whether stimuli are presented on the same trial or on consecutive trials; the changes to the weights will be essentially the same. As in the

Experiment 5, the current simulation points to the limitations of the current trial-based model and the necessity of constructing a real-time version. Adjustable inter-trial intervals, combined with decaying stimulus representations over time, should have an influence on the dynamics of the weight changes. In addition, it would be interesting to explore how interference between stimuli would affect the simulations.

Experiment 7: Extradimensional versus Intradimensional Shifts

Explicit comparison of performance on transfer tests consisting of intra- versus extra-dimensional stimulus shifts can show more transfer across the ID shift than the ED shift. Mackintosh (1974) suggests that this is compelling evidence for shifts in attention to the aspects of the stimuli that are critical to solution of the initial problem. In this experiment, I investigated whether the current model could reproduce the phenomena associated with ED and ID shifts without invoking the dimensional attention explanation. I predicted that the stimulus representations on the competitive layer would become further separated along the "color" dimension in the model than the "orientation" dimension. Thus, during the transfer test, the winning units belonging to networks subjected to an intradimensional shift would be better separated than those of the networks subjected to an extradimensional shift along the dimension critical to the discrimination, resulting in the networks in the intradimensional group learning the transfer test more quickly.

Methods

In this experiment, as in Experiment 6, two dimensions were explicitly represented: The *x* axis represented "color" and the *y* axis represented "orientation". Using this representation allowed for simulation of Mackintosh's original (Mackintosh & Little, 1969) experiment. In order realistically to model the continuous nature of the dimensions being simulated, in this experiment the input

representations were "wrapped". That is, the value <1,0> was considered to be equidistant from <1,1> and <1,10>.

In this simulation, two groups of 25 networks each, group ID and group ED, were initialized. Group ID was first trained to criterion on a discrimination in which a red, horizontal line <8,8> or a red, vertical line <8,4> was reinforced and a green, horizontal line <4,8> or a green, vertical line <4,4> was not. Group ED was first trained to criterion on a discrimination in which a red <8,4> or green <4,4> vertical line was reinforced and a red <8,8> or green <4,8> horizontal line was not. Both groups were then tested on a transfer test in which the relevant dimension was color: blue left-diagonal lines <2,6> and blue right-diagonal lines <2,2> were reinforced and yellow left-diagonal <6,2> and yellow right-diagonal <6,6> lines were not.

Results and Discussion

As can be seen in Figure 4-7, there was no difference between the ED and the ID groups on either the initial training or on the transfer test. Thus, of all of the data that have been explored in this chapter, only the ED versus ID effect does not seem to be explainable in terms of differentiation alone.

If one looks at the positions of the winning units during the transfer tests (Figure 4-8 and Figure 4-9) , however, it seems that the networks did behave as expected with respect to differentiation. As a result of changes in the competitive layer that occurred during initial training, the winning units during the intradimensional transfer test were better separated than those of the extradimensional transfer test. Furthermore, in the case of the ID shift there is no overlap in the neighborhoods of any of the winning units that have opposing reinforcement parameters. In contrast, the winning units of the ED shift network are closer together along the "color" (*x* axis) dimension and there is some overlap between non-reinforced and reinforced areas of activation. This suggests that there should be at least a slight superiority of the ID shifted networks over the ED shifted networks on the transfer test. If anything, however, the nonsignificant trend is in the opposite direction.

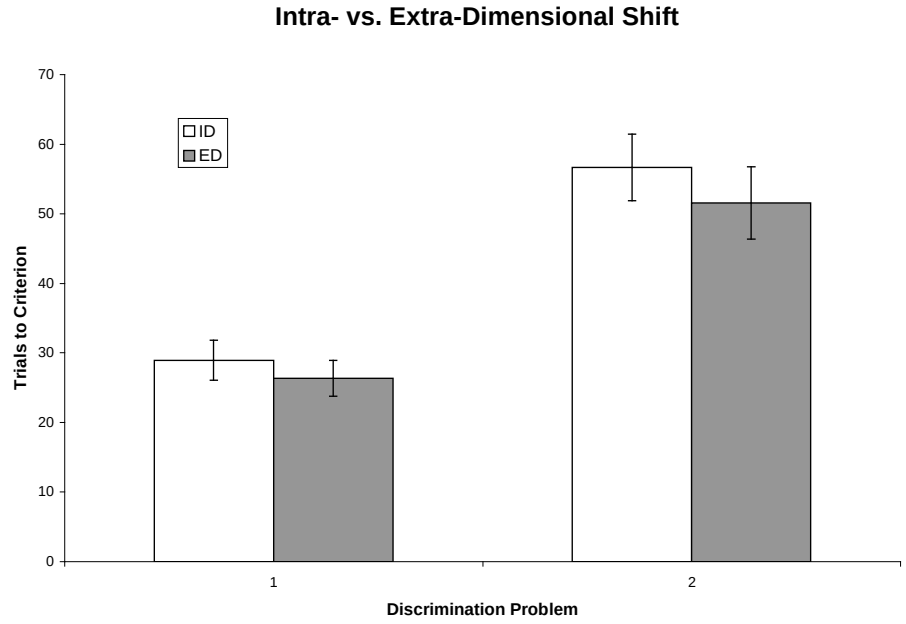


Figure 4-7: Performance on a an initial discrimination and a subsequent transfer test. Discrimination 1 for Group ID consisted of a discrimination in which a red, horizontal line <8,8> or a red, vertical line <8,4> was reinforced and a green, horizontal line <4,8> or a green, vertical line <4,4> was not. Discrimination 1 for Group ED consisted of a discrimination in which a red <8,4> or green <4,4> vertical line was reinforced and a red <8,8> or green <4,8> horizontal line was not. Discrimination 2 for both groups was a transfer test in which the relevant dimension was color: blue left-diagonal lines <2,6> and blue right-diagonal lines <2,2> were reinforced and yellow left-diagonal <6,2> and yellow right-diagonal <6,6> lines were not.

The reason for this unsuspected effect becomes clearer after looking closely at the positions of the stimulus units in initial training, shown in Figure 4-10 and Figure 4-11. For the ID shift group, there is a good deal of overlap between reinforced training units and nonreinforced test units (and vice-versa), suggesting that there will be a large negative effect of associative transfer on the transfer test. For the ED shift group, on the other hand, there is some overlap between reinforced training units and nonreinforced test units, but a greater degree of overlap between reinforced training and reinforced transfer units, suggesting that there will be an overall positive effect of associative transfer on the transfer test. Thus, it appears that in this particular simulation, the effect of associative transfer outweighs the effect of differentiation.

The above explanation for the current result points out a critical problem with a differentiation-based explanation of the ED/ID shift effect. It might on first glance seem that the current differentiation model would produce the appropriate simulation data if one simply further separated the training and transfer input stimulus patterns in order to eliminate the possibility of associative transfer. After all, researchers investigating this issue in animal subjects take pains to eliminate associative transfer as much as possible. Unfortunately, however, the proximity of these input stimuli is exactly what would allow the differentiation approach to work at all in this case. If the inputs were separated such that there was no interaction between the winning units on the competitive layer, then there would be little difference between the ID trained and the ED trained network's competitive layers, and a consequent lack of the ED/ID shift effect.

In sum, it seems that a differentiation model can account for many effects that have been attributed to dimensional attention, but the ED/ID shift transfer effect is not one of them. The implications of this are discussed in the following section.

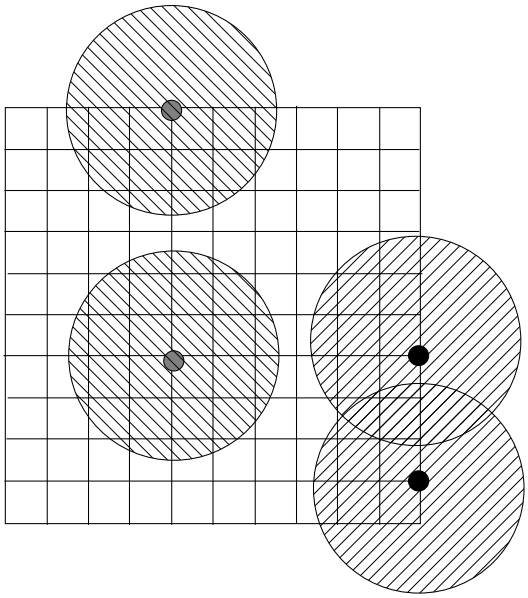


Figure 4-8: The positions of the winning competitive units after 50 trials of the intradimensional transfer test. Black units are reinforced, gray units are nonreinforced. The larger circles around the central point denote the Gaussian area of activation around the winners.

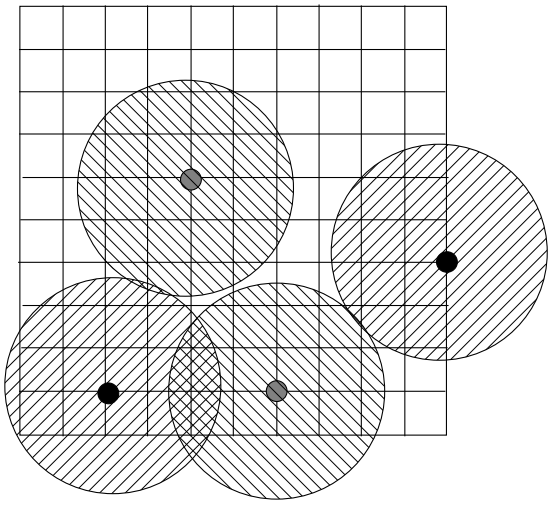


Figure 4-9: The positions of the winning competitive units after 50 trials of the extradimensional transfer test. Black units are reinforced, gray units are nonreinforced. The larger circles around the central point denote the Gaussian area of activation around the winners.

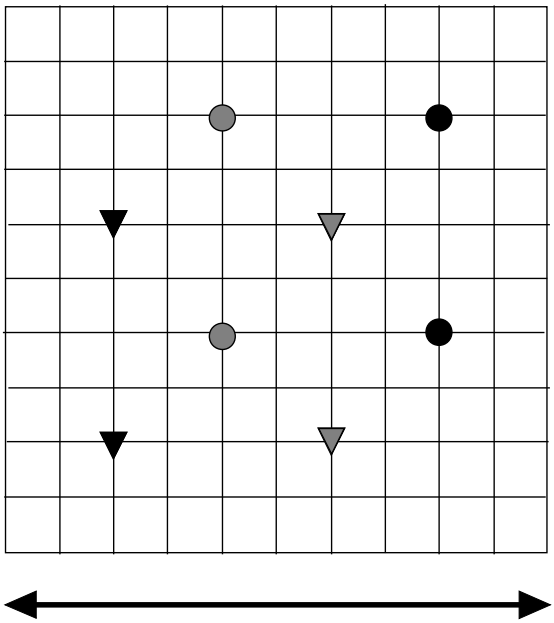


Figure 4-10: Stimulus input winning units for intradimensional training (circles) and the transfer test (triangles). Black units are reinforced, gray units are nonreinforced. There is a good deal of overlap between reinforced training units and nonreinforced test units (and vice-versa), suggesting that there will be a large negative effect of associative transfer on the transfer test. The arrow depicts the axis along which the competitive layer stretches.

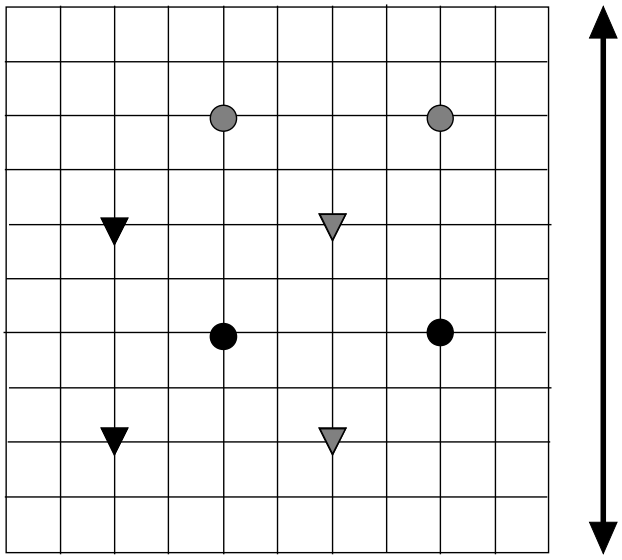


Figure 4-11: Stimulus input winning units for extradimensional training (circles) and the transfer test (triangles). Black units are reinforced, gray units are nonreinforced. There is some overlap between reinforced training units and nonreinforced test units, but a greater degree of overlap between reinforced training and reinforced transfer units, suggesting that there will be an overall positive effect of associative transfer on the transfer test. The arrow depicts the axis along which the competitive layer stretches.

General Discussion

The current chapter demonstrates that several effects that have been attributed to the concept of dimensional attention in fact may be explained simply as a result of differentiation processes operating on topographically organized stimulus representations. It was shown that the overtraining reversal effect , transfer along a continuum, and reversed transfer along a continuum can all be explained as a result of differential differentiation of the discriminanda. Data pertaining to the influence of stimulus comparison were also shown to be accounted for by the current model. Thus, it seems that a simple differentiation mechanism, when well-defined, is perhaps more general and powerful than generally considered: in addition to preexposure effects, differentiation can also account for several phenomena that are usually attributed to higher-level processes.

The ED/ID effect

Whereas differentiation was able to account for several phenomena, simulations of ED/ID transfer experiments demonstrated a complete lack of that effect. There are two possible reasons for this. First, differentiation may simply be inappropriate as a mechanism for perceptual learning. However, a second possibility is that the ED/ID shift is a result of a higher level, "top-down" system similar to Mackintosh's dimensional attention analyzers. That dimensional attention requires such a higher-level process is supported by at least two lines of evidence. First, whereas the ED/ID effect is somewhat difficult to obtain in "lower" species such as rats and pigeons, with some authors obtaining such effects (Klosterhalfen, Fischer, & Bitterman, 1978; Mackintosh & Little, 1969) and others not (Couvillon, Tennant, & Bitterman, 1967; Hall & Channell, 1985; Tennant & Bitterman, 1973), in "higher" species such as monkeys and humans the effect is more robust (Roberts, Robbins, & Everitt, 1988). Second, Dias, Roberts, and Robbins (Dias, Robbins, & Roberts, 1996) have provided evidence from the monkey that the prefrontal cortex is the critical region for dimensional attentional set-shifting. These two pieces of evidence suggest

that dimensional attention is a higher-level process that is most developed in higher species, and is subserved by higher-level brain regions that appear late in evolution. Thus the current model appears to be successful at simulating effects that are readily observed in lower species, but is less able to handle higher-order functions characteristic of higher species. For this, higher-order top-down processing would need to be added to the network. In order to advance the current model to produce these effects, it may be more feasible not to look for an alternate bottom-up process, but instead to incorporate a high level attentional process.

While the fact that ED vs. ID transfer was not replicated in this chapter may point to one reason that the phenomenon is not able to be modeled using a simple differentiation process, the fact that differences in ED vs. ID transfer do occur reliably in certain species suggests that the phenomenon is real. Since this effect cannot be accounted for by the current model, whereas other data attributed to dimensional attention can, it would be useful at this point to try to understand what the critical differences are between those experiments that can, and those that cannot, be accounted for by differentiation.

Real-Time Effects

The reason that the current model was unable to account for two sets of experimental data investigated in this chapter is the fact that the model is currently trial-based. As such, it cannot be expected to simulate effects that are critically dependent on intra- or intertrial temporal parameters. It seems clear, however, that if a simple representation of time were incorporated into the current model then it would be able to account for these data. For example, consider an alteration of the model such that each trial and inter-trial interval is divided into, say, one second bins. In addition, each stimulus would be present for only the first second of the trial. In the subsequent absence of the stimulus, the activity of the input units would be set to zero and the activity of the competitive units would decay exponentially. If this real-

time version of the model were run, I would expect it to account for the data presented in Experiments 5 and 6 as follows.

In Experiment 5, the effect of mixed versus blocked preexposure on easily discriminable stimuli was investigated. Honey et al. (1996) have shown that with this type of stimuli blocked exposure tends to facilitate subsequent discrimination learning more so than mixed exposure. A simple explanation for this finding, in the real-time version of the model, would be that during mixed exposure there is a greater degree of interference between the stimuli. As a result, the stimulus representations on the competitive layer would not be as well-separated as in the case of networks subjected to blocked exposure, and the blocked exposure networks would perform better on subsequent tests of discrimination.

In Experiment 6, the influence of stimulus comparison was investigated. Saldanha and Bitterman (1951) showed that the opportunity to compare stimuli on a trial led to a facilitation of subsequent discrimination. In the simulations of this experiment the appropriate pattern of results was not produced because there was essentially no difference between paired and unpaired stimulus presentations in the way in which weights were changed. A real-time version of the model, however, would allow for the manipulation of the sampling rate. In the current model, the network samples the input once at the beginning of each trial. In a real-time version it would be possible to allow the network to sample the stimuli a number of times during each trial (e.g., once per second). One would have to make the additional assumption of the existence of some kind of confidence criterion that guides the animal in making its choice of stimuli (i.e., the animal repeatedly samples a stimulus until it surpasses an internal confidence threshold regarding the identity of the stimulus). Thus, on easier discriminations, such as those in which the stimuli vary along different dimensions, one would expect the confidence threshold to be reached more quickly. As a result, animals in the more difficult paired condition would spend more cycles per trial sampling the stimuli. If one assumes that the differentiation mechanism updates on each perceptual sample then one would expect

more differentiation to occur for the animals in the paired group than for those in the unpaired group, thereby facilitating their performance. This mechanism makes the plausible prediction that although animals in the paired condition may require the same number of trials to learn the discrimination, they should take longer to respond.

Summary

In sum, the current chapter demonstrated that several results typically ascribed to mechanisms of dimensional attention can in fact be accounted for by a simple differentiation mechanism. As differentiation is a lower-level process than dimensional attention, and as differentiation can account for many additional effects such as those involving preexposure, differentiation may provide a more parsimonious account of the data discussed in this chapter. Furthermore, the development of a real-time version of this differentiation model would allow it to account for a larger body of data, including those presented in Experiments 5 and 6. Finally, the current chapter demonstrated that although differentiation may account for some of the data attributed to dimensional attention, it cannot produce the ED/ID shift pattern of effects. Focusing on the differences between those data that are explicable in terms of differentiation and those that are not may provide additional insight into the nature of dimensional attention.

CHAPTER 5 ACQUIRED EQUIVALENCE AND ACQUIRED DISTINCTIVENESS

Introduction

A second set of effects, separate from those addressed in dimensional attention theories, that have been attributed to attentional processing involves what appears to be the influence of task-specific information on discrimination learning. If two stimuli share the same reinforcement history, labels, or associated responses their discriminability may be reduced, thereby leading to a phenomenon called "acquired equivalence" (AE). On the other hand, if the same two stimuli have vastly different reinforcement histories, labels, or associated responses they may become more discriminable, in which case they are said to have undergone a process of "acquired distinctiveness" (AD). Like dimensional attention, these effects are usually explored through a transfer test in which subjects are trained to categorize objects and then are tested to see whether the objects are then more (or less) quickly discriminated.

The critical aspect of these phenomena that separates them from other issues addressed in this dissertation is that they require information additional to visual feature representations that can be refined to make similar features discriminable. As will be seen in the following section, the key commonality in experiments on AD and AE is that they introduce task-specific information that is not necessarily directly related to the discriminanda (i.e., may be of a different modality etc), but that allows the discriminanda to be distinguished not on the basis of stimulus similarity per se, but on the basis of some kind of task-specific semantic similarity.

The model as developed up to this point is capable of categorizing inputs on the basis of their physical similarity only. With experience, the model is able to refine its

representation of the degree of similarity of stimuli, and as such can learn to discriminate initially indistinguishable stimuli. However, the model has no ability to deal with task-specific constraints that are not in line with the stimulus representations. In this chapter I explore a mechanism for incorporating an "event" signal into the competitive learning architecture that allows the model to classify dissimilar inputs to similar outputs, thus producing effects of AD and AE.

In the following section of this chapter, I briefly review evidence for the phenomena of acquired equivalence and distinctiveness. I then discuss theories which have attempted to provide mechanisms for these phenomena. Finally, I present a version of the current model that can account for these effects, based on the assumption that reinforcement or labeling information can become incorporated with the stimulus representation to form an event representation. Next I present simulations of AD and AE effects and discuss the implications of the results.

Evidence for Acquired Equivalence and Distinctiveness

The first work that explicitly investigated AD was Lawrence's (1949) experiment that was described at the beginning of Chapter 4. To summarize, rats were trained on a simultaneous black-white discrimination and then were tested on a transfer test during which either both ends of the maze were black and response to the right was rewarded or both ends were white and response to the left was rewarded. Animals in the latter experimental condition learned the successive discrimination more readily than controls, suggesting that the black and white stimuli had "acquired distinctiveness". As discussed previously, Lawrence attributed this effect to dimensional attention. Other non-attention based explanations, in addition to the idea presented in the previous chapter, have been proposed, as will be seen in the following section.

Additional evidence for AD, which focuses on the critical contribution of categorization, has also been produced. For example, Norcross and Spiker (1957) found that children given training in applying different names to two faces were

superior in subsequent test performance to controls who were pretrained in a same-different task. This result was confirmed by Spiker and Norcross (Spiker & Norcross, 1962). The authors suggest that training in which critical stimuli become linked to different events generates transfer to subsequent discrimination learning. However, their result may have been due to a difference in the difficulty of two stages. Reese (1972) addressed this issue by conducting an experiment in which children were trained with three nonsense figures, learning to apply one verbal label to two of them and a different label to the third. All groups received the same training thus there were no differences in difficulty between groups. Children were subsequently divided into two test groups and received successive discriminations in which they were instructed to press one button in response to one of the pretrained cues and a different button to another pretrained cue. For children in the first test group—the acquired distinctiveness group—the two stimuli had had different labels during pretraining. For children in the second group—acquired equivalence—the two stimuli had had the same labels during pretraining. The researchers found that subjects in the distinctiveness condition learned the discrimination more quickly. Norcross (1958) reported similar results.

Grice and colleagues (1958; Grice & Davis, 1960; 1963) have obtained evidence for AD and AE in adults. In a preliminary training stage subjects were asked to press one key in response to a given tone and to press a different key when a light or a different tone was presented. In the next stage the light was presented alone and was reinforced with an air puff to the subject's eye, leading to the acquisition of a conditioned eye blink in response to the light. In the final phase generalization from the light to the tones was measured and it was found that there was significantly more generalization from the light to the tone which had been pretrained with the same response.

Honey and Hall (1989) have investigated the influence of an initial phase of discrimination training on a generalization test in rats. In order to distinguish the CR of the test stage from the first stage, the nature of reinforcer was changed from

appetitive to aversive. Two groups of rats received Pavlovian training with two auditory stimuli, A (a burst of white noise) and B (a tone). For the first group both sounds signaled food, while for the second group B was reinforced and A was not. The effect of this preliminary training on generalization between A and B was tested by establishing A as a CS for a new response and then testing for B's ability to evoke that response. A was next trained as a signal for shock in both groups. Little generalized suppression was shown for the group given initial discrimination training suggesting that these subjects were better at discriminating A from B. Since there was no guarantee, however, that the animals acquired the same degree of aversion to A, Honey and Hall performed another control experiment. Two groups of rats received initial training with three auditory stimuli A, B, and C. For both groups, B (a tone) and C (a clicker) had different consequences: one was followed by food and the other was not. For the first group, A (white noise) was reinforced whereas for the second group A was not reinforced. All subjects then received A→shock pairings followed by a generalization test to B and C. Each group showed more suppression to the stimulus that had had stage one training equivalent to A. The advantage of this procedure is that the degree of suppression does not depend on whether it was reinforced appetitively during the first stage.

Finally, a phenomenon called the "differential outcomes effect" provides additional evidence for both AD and AE (Peterson & Trapold, 1980). Rats were tested on a food-rewarded (conditional go left/go right) discrimination between a tone and a clicker. Those in the correlated condition experienced a light only after a correct response in the presence of one cue; those in the uncorrelated condition received a light after 50% of rewarded responses whether paired with the tone or the clicker. The researchers found that accurate choice performance increased more quickly in the correlated group. This is called the differential outcomes effect because the superior performance of the correlated group depends on the fact that the outcome of the correct response is reliably different for the two trial types. Similar results are produced when the reinforcers differ in their nature (Trapold, 1970), in their timing (Carlson & Wielkiewicz, 1972) or magnitude (Carlson &

Wielkiewicz, 1976). The differential outcomes effect has been interpreted as showing that expectancies can act as effective cues in the control of choice responding.

The above evidence suggests that discrimination training, in which each of a pair of cues becomes associated with a different outcome, enhances performance on a further discrimination task with the same cues. However, as Hall (1991) notes, none of the above studies can tell us whether this is due to acquired equivalence, acquired distinctiveness, or both. Goldstone (1994, Experiment 2), however, has investigated this particular issue in the context of a human categorization task. In this experiment, stimuli consisted of sixteen squares, each of which had a brightness value of 1, 2, 3, or 4 and a size value of 1, 2, 3, or 4. Subjects were given different categorization rules: for one group, size was relevant (objects of size 1 or 2 belong to category A, those of size 3 or 4 belong to B), for a second group brightness was relevant (objects of brightness 1 or 2 belong to category A, the others to B), and a third group received no categorization training. The results of a subsequent same-different test in which the trained stimuli were used showed that perceptual discriminations along the categorization-relevant dimension were better for both categorization groups than for controls, thus providing specific evidence for acquired distinctiveness. The evidence for acquired equivalence, however, was mixed. Two types of acquired equivalence were possible: acquired equivalence of the irrelevant dimension or acquired equivalence within the relevant dimension. In the first case, the brightness group was significantly worse than controls at making size discriminations; there was no difference, however, between the size group and controls on brightness discriminations. Thus the evidence regarding acquired equivalence of the irrelevant dimension was ambiguous. In the second case, both experimental groups showed that squares with different values along the trained dimension were more discriminable, demonstrating a trend in the opposite direction of acquired equivalence.

The evidence against acquired equivalence within the categorization-relevant dimension is also evidence for a dimension-wide shift in attention. For example, Goldstone found that when squares of size 1 and 2 were categorized as A and when squares of size 3 and 4 were categorized as B, there was sensitization to the difference between sizes 1 and 2 relative to controls. According to Goldstone, this suggests that subjects learn to attend not just to specific values on a given dimension, but to the dimension itself. However, although the difference between sizes 1 and 2 was sensitized, the difference between 2 and 3—a difference critical to correct categorization—was much further sensitized. So from this experiment it appears that enhanced attention to local areas of a dimension are also possible.

Theoretical Interpretations

According to Hull (1939), primary generalization occurs between stimuli that are similar, due to common features. Secondary or learned similarities, however, can mediate generalization between otherwise dissimilar stimuli. Hull's account of the mechanism underlying this is framed in the S→R theory that was popular at the time. If two stimuli, SA and SB, are followed by the same response RX then each of them will develop associative links with it (SA→RX and SB→RX). When RX occurs, it will elicit a set of feedback cues SX that will be evoked upon presentation of either SA or SB. Training subjects to then make a new response, RY, in response to SX, allows for the formation of SX→RY associations. Since SB is capable of evoking SX, however, it too will be able to elicit RY even though the association has not been trained directly. From a modern associative theory perspective, mediation theory does not have to be strictly S→R: representations of SA and SB could be linked directly to the representation of SX. While mediation theory provides a mechanism for acquired equivalence, Hull does not attempt to deal with acquired distinctiveness. In order for mediation theory to account for acquired distinctiveness it must include a processes that reduces the role of common elements in producing generalization from A to B. It is not clear how associations between SA and SX and SY and SB would be able to do this.

Contrary to mediation theory, the Gibsons' theory of differentiation, described in Chapter 3, ignores the possibility of acquired equivalence. Furthermore, differentiation does not specify why explicit training enhances the discrimination effect. In other words, although differentiation does provide a mechanism for acquired distinctiveness, the mechanism is not dependent on reinforcement. As a result, their theory may provide a partial mechanism for acquired distinctiveness but it cannot account for the fact that differential previous reinforcement significantly facilitates discrimination.

A third account of acquired equivalence and distinctiveness, although not usually specified as such, can be found in backpropagation based neural network models. This is exemplified by the concepts of "redundancy compression" and "predictive differentiation" proposed by Gluck and Myers (1993) to be critical functions of the hippocampal region. Their model consists of an autoencoder which, given input representing the stimuli present, learns to reconstruct the inputs as well as to predict future reinforcement. The critical aspect of the model is that the hidden layer is narrower than the input and output layers. As such, the network is forced to develop hidden layer representations that compress redundancies in the input (as in acquired equivalence) while at the same time preserving predictive information (as in acquired distinctiveness). The network is able to develop this compressed stimulus representation through the backpropagation of the error signal from the output units to the input units in order appropriately to adjust the weights in each layer of the network (see Chapter 2).

The evidence for acquired distinctiveness makes it clear that differential categorization of stimuli facilitates their discrimination. Differentiation theory, however, does not address the importance of reinforcement in this result. The data regarding acquired equivalence are quite a bit less conclusive, so while mediation theory may provide an adequate account of this phenomenon, it is difficult to see how it could explain the more robust acquired distinctiveness effect. Backpropagation, while providing a reasonable simulation of AD and AE effects,

requires full supervision rather than just a reinforcement signal and thus is not compatible with the current Kohonen-based model of differentiation and perceptual learning. Thus, in the following section I develop a mechanism for producing AD and AE effects that is simple and compatible with the current model of differentiation and perceptual learning.

Adding Task-Specific Information to the Model

The current model bases discrimination solely on input similarity and as such has no mechanism for dealing with task-dependent effects. One method for achieving classification based on task-dependent similarity rather than input similarity is to provide the competitive learning mechanism with more information. Machine learning researchers have done this with a "teaching" input—corresponding to reinforcement or a categorization label—that tells the network which inputs should be mapped to the same output. The general effect of the incorporation of this information is to move competitive-layer input representations with similar reinforcement histories closer together and representations of inputs with differential reinforcement histories apart (de Sa, 1994; Rumelhart & Zipser, 1986).

The incorporation of reinforcement information with stimulus representations is a mechanism that seems likely to lead to effects such as AD and AE, and is easily compatible with the current model. Furthermore, two disparate pieces of evidence from the animal learning field support the likelihood that this may occur. First, Bouton and colleagues have provided much evidence suggesting that separate representations of stimuli may be maintained according to context including reinforcement history (e.g., spontaneous recovery, reinstatement, renewal, see Bouton, 1994). Second, "expectancy theory", the idea that the representation of the reinforcer is encoded in an associative structure which mediates the CR, has been a subject of much supporting research (see Dickinson, 1989).

In the extended model, it is assumed that the incorporation of external information is not limited to reinforcement signals; any type of cue that co-occurs with a stimulus—response, verbal label, auditory signal—is a candidate for inclusion with the stimulus representation. Furthermore, multiple cues may separately be linked with a given stimulus representation to form an "event" representation. Like clusters of bits of stimulus information in configural learning theory (e.g., Pearce, 1994), these potentially large sets of cues may be represented as configurations in a multidimensional psychological space, and the event representations that are retrieved from memory will be those that best match external cues present at the time of retrieval. In the following section I describe more specifically how the model incorporates reinforcement information with stimulus representations.

Architecture of the Extended Model

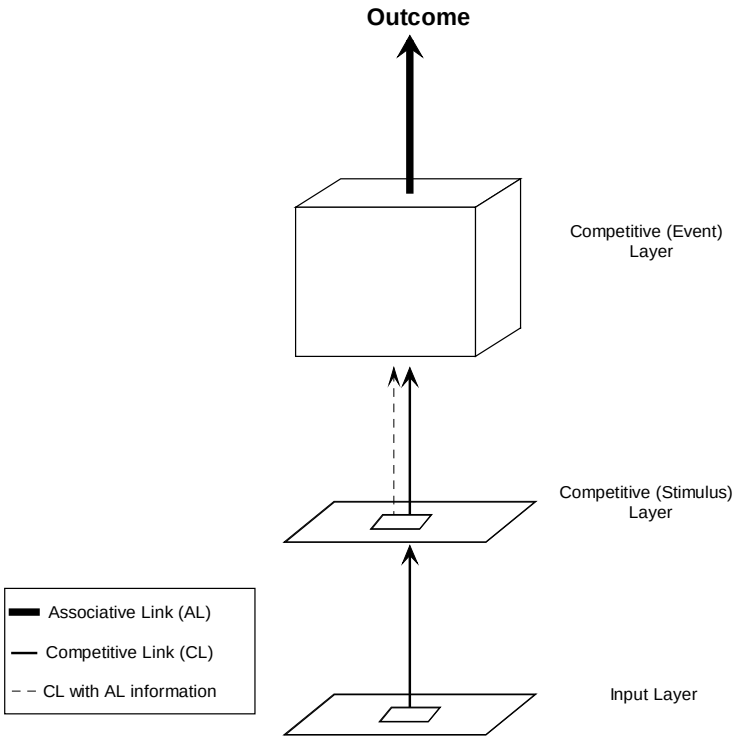


Figure 5-1: The differentiation model extended to incorporate reinforcement information with stimulus representations. As in the original differentiation model, input of a stimulus to the network leads to the activation of a two-dimensional layer of input units and these inputs feed into a two-dimensional layer of competitive units to form the stimulus representation. The

difference is that in the extended model, the stimulus representation units are fully connected to an additional, three-dimensional layer of "event" units. The three dimensions of the event layer consist of the two stimulus dimensions plus a third dimension representing the predicted outcome of a stimulus presented to the network. Thus an event (i.e., stimulus representation plus associated outcome information) is represented as a point in this topographically organized three-dimensional space.

Figure 5-1 illustrates the structure of the extended network. The learning mechanisms in the extended network are identical to those used in the original model; the only differences are those of architecture. As in the original differentiation model, input of a stimulus to the network leads to the activation of a two-dimensional layer of input units and these inputs feed into a two-dimensional layer of competitive units to form the stimulus representation. In the extended model, these stimulus representation units are fully connected to an additional, three-dimensional layer of "event" units. The three dimensions of the event layer consist of the two stimulus dimensions plus a third dimension representing the predicted outcome of the presentation of a given stimulus to the network. Thus an event (i.e., stimulus representation plus associated outcome information) is represented as a point in this topographically organized three-dimensional space. The proximity of two event representations is therefore dependent on both stimulus similarity and reinforcement history. Two similar stimuli with similar reinforcement histories will remain close together along the predicted outcome dimension, resulting in a relatively large degree of generalization between them. If two stimuli are presented and differentially reinforced, however, then winning units in the event layer will tend to become increasingly far apart as the CR is developed, leading to the effect of acquired distinctiveness.

Operation of the Extended Model

On a training trial, the x and y values of a stimulus **S** are fed into the network and the unit in the input layer whose coordinates correspond to the x and y values of **S** becomes maximally activated. This unit affects nearby units such that those that are within a certain radius (σ) of the unit will also be activated proportionally to their distance from it, thereby forming a "blob" of activation on the input layer.

Equation 5-1 is the input layer activation function for a stimulus whose input layer location is $i^*=\langle x,y\rangle$. The degree of activation of a input layer element with location i_j is 1 for $i_j=i^*$ and decreases as a function of the distance between i_j and i^*. The width of the Gaussian is σ^2.

$$a_j = \exp\left(-\frac{\left\|i_j - i^*\right\|^2}{2\sigma^2}\right) \tag{5-1}$$

Every unit in the input layer is connected to each unit in the subsequent competitive layer **C**. The link between one input unit j and one competitive unit k is called w_{jk} Thus for competitive unit k there exists a corresponding set of weighted links, $\mathbf{w}_k$, which consists of the set of links between each unit in the input layer and unit k.

In order that the competitive layer is roughly topographic from the start, w_{jk} are initialized as Gaussian distributions, centered over the corresponding input unit coordinates, that are corrupted by a significant degree of uniformly distributed random noise r.

$$w_{jk}(0) = \exp\left(-\frac{\left\|i_j - c_k\right\|^2}{2\sigma^2}\right) + r(-0.3,0.3) \tag{5-2}$$

In the above equation, i_j is an $\langle x,y\rangle$ location in the input layer and c_k is an $\langle x,y\rangle$ location in the competitive layer.

When a stimulus is presented to the network, the Euclidean distance between $\mathbf{w}_k$ and the pattern of activation on the input layer (**a**) are calculated (see Equation 5-4) and compared (see Equation 5-3) for each competitive unit.

$$k^* = \arg\min(\mathbf{d}) \tag{5-3}$$

$$d_k = \sqrt{\sum_j (a_j - w_{jk})^2} \tag{5-4}$$

As in the input layer, the winning unit influences neighboring units, and they are activated proportionally to their distance from the winner. Competitive unit activation a_k is 1 for $c_k = c^*$ and falls off with distance $\left\| c_k - c^* \right\|$ where c_k and c^* are the grid coordinates of units k and k^* respectively. As a result, units close to the winner are significantly more active than units further away. The degree of activation of a unit k is determined by a Gaussian defined as follows:

$$a_k = \exp\left(-\frac{\left\| c_k - c^* \right\|}{2\sigma^2} \right) \tag{5-5}$$

As mentioned above, the extent of influence, or neighborhood size, of the winning unit is reflected in the Gaussian width parameter σ.

In addition to the winner, the $\mathbf{w}_k$ corresponding to a set of nearby units are also updated in proportion to their proximity to the winner. The result is that after training is complete, nearby competitive units respond to nearby input patterns. The effect of this is to set up regions of the competitive layer which code for similar representations, that is, those with common features. Equation 5-6 is responsible for the competitive weight update process.

$$\Delta w_{jk} = \delta_k \cdot a_k \cdot \left(a_j - w_{jk} \right) \tag{5-6}$$

The competitive learning rate, δ_k, decreases as the outcome of trials become well predicted (see Equation 5-10).

Each stimulus unit has an associative link to the outcome representation. After the stimulus layer is activated, the predicted outcome of that pattern of activation is computed as the sum of each unit's activation multiplied by its associative weight (see Equation 5-7).

$$O_1 = \sum_k V_k \cdot a_k \tag{5-7}$$

The stimulus layer is also fully connected to the three-dimensional event layer via links w_{kl}. The three dimensions of the event layer consist of the two stimulus dimensions plus a third dimension representing the CR, or predicted outcome (O_1), of the network.[6] Input to the event layer consists of the pattern of activation of the stimulus layer plus the outcome (O_1) predicted by the stimulus layer.

The event layer is fully connected to the actual outcome representation (O_2), and the associative strengths of these links, as well as the stimulus layer associative links, are updated on each trial by the Rescorla-Wagner rule (Equations 5-9 and 5-10). The output of the network, or CR, consists of the sum of the associative strengths of the event competitive units scaled by their activation (O_2; see Equation 5-8).

$$O_2 = \sum_l V_l \cdot a_l \tag{5-8}$$

$$\Delta V_k = \alpha_k \cdot \beta \cdot (\lambda - O_2) \tag{5-9}$$

$$\Delta V_l = \alpha_l \cdot \beta \cdot (\lambda - O_2) \tag{5-10}$$

General Methods

Identical network parameters were used for all simulations presented in this chapter. The network consisted of ten by ten grids of perceptual and stimulus competitive units and a ten by ten by ten block of event competitive units. The

[6] As in the original model, the CR varies between 0 and 1. However, in the event layer, the CR dimension ranges from 0 to 10 in order to be proportional to the stimulus representations. The actual CR is mapped onto a position on the CR dimension by multiplying it by 10.

specific parameters were as follows: $\alpha(0)=1$, $\beta=0.005$, $\delta(0)=0.05$, $\gamma=0.95$, $\lambda=1$ (reinforced trials) or 0 (nonreinforced trials), $\sigma^2(0)=2.0$.

The data of interest in this section consist of the degree to which the network will generalize "perceptually" between stimuli after different types of training. Thus, instead of looking at the CR, I look directly at the degree to which the winning unit for a given stimulus is activated in the presence of another stimulus. For example, to determine the degree of generalization between stimulus A and stimulus B I look at the activation of the winning event competitive unit for B in the presence of A, and the activation of the winning competitive unit for B in the presence of A. In all of the simulations in this chapter, generalization between stimuli in a pairwise discrimination is calculated as the average of these two values and will hereafter be referred to simply as "generalization". Since the activation refers to Euclidean distance, a larger activation reflects *less* generalization and a smaller activation reflects *more* generalization.

Experiment 1: Labeling Stimuli

Reese (1972) looked at the effect of categorization training on subsequent discrimination by conducting an experiment in which children were trained with three nonsense figures, learning to apply one verbal label to two of them and a different label to the third. All groups received the same training thus there were no differences in difficulty between groups. Children were subsequently divided into two test groups and received successive discriminations in which they were instructed to press one button in response to one of the pretrained cues and a different button to another pretrained cue. For children in the first test group—the acquired distinctiveness group—the two stimuli had had different labels during pretraining. For children in the second group—acquired equivalence—the two stimuli had had the same labels during pretraining. The researchers found that subjects in the distinctiveness condition learned the discrimination more quickly. In

this simulation, I investigate whether the extended version of the differentiation model can produce a similar effect.

Methods

In this simulation, one group of 10 networks was initialized and then trained on the following serial discrimination: A<5,5>+, B<3,3>+, C<7,7>-, in which the positive reinforcement signal served as the stimulus label. Note that A is equidistant from B and C. Subsequent to attainment of the criterion of 90% correct on 3 consecutive blocks of 10 trials, learning was turned off and the networks were tested on one trial of each of two simultaneous discriminations: A vs. C and A vs. B.

Results and Discussion

As can be seen in Figure 5-2, there was less overlap between A and C (distinctiveness condition) than there was between A and B (equivalence condition). Although A was equidistant from B and C in terms of perceptual similarity alone, differential reinforcement of A and C caused the winning units for each stimulus to be further separated along the predicted outcome dimension than were the winning units for A and B. The success of this simulation demonstrates that the extended model is able to produce the basic acquired distinctiveness effect.

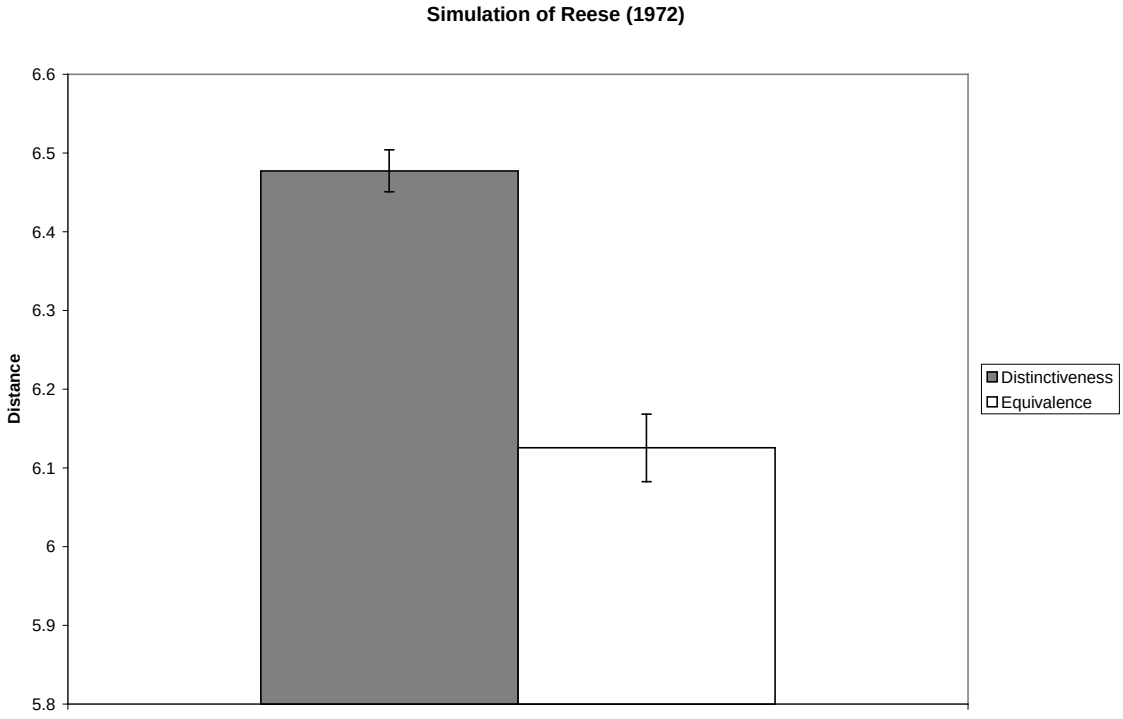


Figure 5-2: Discrimination performance after categorization training. Networks were initially trained on the following discrimination A<5,5>+, B<3,3>+, C<7,7>- (not shown here). Next, the networks were tested in extinction on one trial of each of two simultaneous discriminations: A vs. C and A vs. B. Generalization between A and C (distinctiveness condition) was less than generalization between A and B (equivalence condition) even though the stimuli were perceptually equidistant.

Experiment 2: Differential Outcomes Effect

As discussed earlier, the differential outcomes effect provides additional evidence for AD and AE (Peterson & Trapold, 1980). Animals trained in a correlated condition, in which an outcome is delivered only in the presence of one cue, are better able to discriminate cues than animals in an uncorrelated condition in which an outcome is presented 50% of the time after each cue. This is called the differential outcomes effect because the superior performance of the correlated group depends on the fact that the outcome of the correct response is reliably different for the two trial types. In this experiment, I investigated whether the extended model could produce this different type of acquired distinctiveness effect.

Methods

In this simulation, two groups of 10 networks each (Group Differential Outcome (DO) and Group Same Outcome (SO)) were initialized. Group DO was trained on a discrimination between A<5,5> and B<3,3>, in which A was reinforced 100% of the time and B was never reinforced. Group SO was trained for the same number of trials on a discrimination in which A and B were each reinforced 50% of the time. Next, the networks were tested, with learning turned off, on one trial of the same A vs. B discrimination.

Results and Discussion

As can be seen in Figure 5-3, there was less overlap between A and B for the group that had been trained with differential outcomes than for the group that had been trained with the same outcome, demonstrating that the extended model is capable of producing the differential outcome effect. Peterson and Trapold (1980) explained their result by suggesting that expectancy (of the outcome) served as an additional cue controlling choice responding. The current model instantiates this view by considering expectancy to be simply another factor that is combined with stimulus representations in the construction of an event representation. Expectancies are represented topographically, just as stimuli are, thus differences between expectancies contribute to the degree of generalization between two events just as do differences between stimuli.

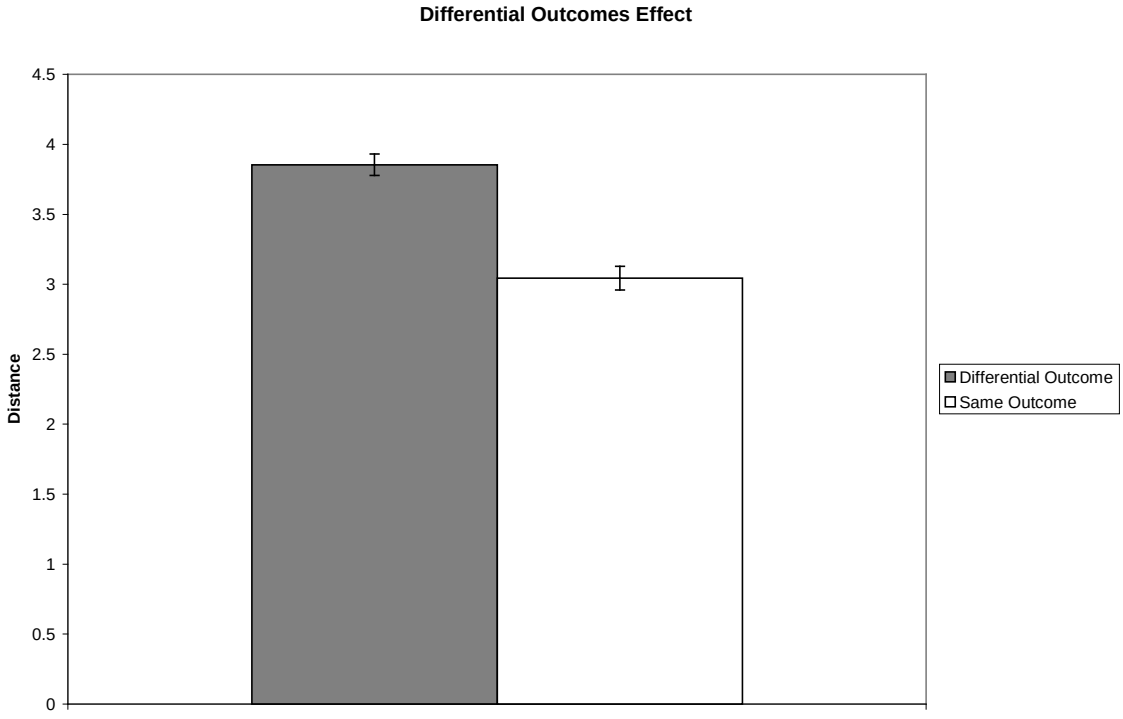


Figure 5-3: Group Differential Outcome was trained on a discrimination between A<5,5> and B<3,3>, in which A was reinforced 100% of the time and B was never reinforced. Group Same Outcome was trained for the same number of trials on a discrimination in which A and B were each reinforced 50% of the time. This graph shows generalization between A and B at the end of this training. Group Same Outcome evidenced more generalization between A and B than did Group Differential Outcome.

Experiment 3: The Influence of Categorization on Perception

In this experiment, I simulate Goldstone's (1994) experiment on the influence of categorization on perception. (The experiment and results are detailed in the introduction to this chapter.) In Goldstone's experiment, stimuli consisted of sixteen squares, each of which had a brightness value of 1, 2, 3, or 4 and a size value of 1, 2, 3, or 4. Subjects were given different categorization rules: for one group, size was relevant (objects of size 1 or 2 belong to category A, those of size 3 or 4 belong to B) for a second group brightness was relevant (objects of brightness 1 or 2 belong to category A, the others to B), and a third group received no categorization

training. This experimental design allowed Goldstone to explore several aspects of AD and AE . First, he looked at whether there was any evidence for AD. Next, he looked at AE of the irrelevant dimension and within the relevant dimension. Finally, he looked at whether or not there was local sensitization within a dimension. Each of these aspects is explored in the current simulation.

Methods

Sixteen stimuli were created which had values of 2, 3, 4, or 5 on dimension 1 and values of 2, 3, 4, or 5 on dimension 2. The coordinates of these stimuli in the input space are shown in Figure 5-4.

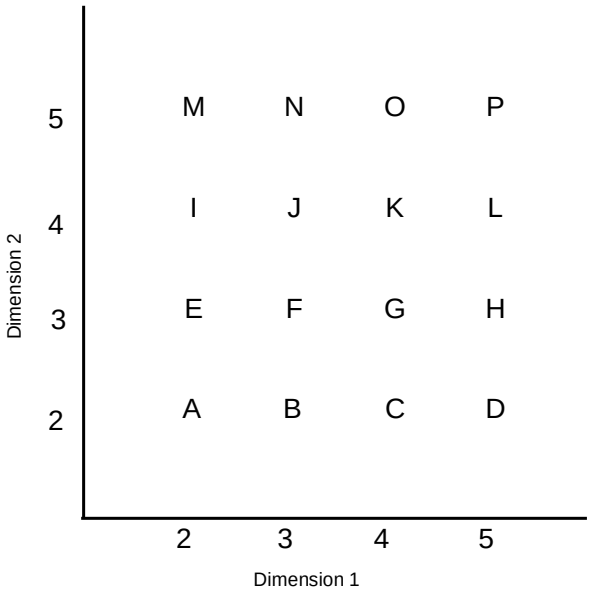


Figure 5-4: Coordinates of the sixteen stimuli used for simulations of Goldstone (1994).

Three sets of 10 networks each were initialized. Group 1 was trained on the following serial discrimination in which categorization was performed along dimension 1: A+ B+ E+ F+ I+ J+ M+ N+ C- D- G- H- K- L- O- P-. Group 2 was trained on a different serial discrimination in which categorization was performed along dimension 2: A+ B+ C+ D+ E+ F+ G+ H+ I- J- K- L- M- N- O- P-. Each network was trained on the appropriate discrimination for 5 blocks of 10 trials each.

The third group of networks, Group Control, was exposed to the stimuli for 5 blocks of 10 trials each with no reinforcement. After discrimination training, each group was presented with 16 pairwise discriminations and, as in the previous experiments in this chapter, the degree of generalization between the two stimuli was determined for each network. The pairwise discriminations, which will be discussed in detail in the following section, were as follows:

1	2	3	4	5	6	7	8
B vs. C	F vs. G	J vs. K	N vs. O	I vs. E	J vs. F	K vs. G	H vs. L

9	10	11	12	13	14	15	16
A vs. B	I vs. J	K vs. L	C vs. D	I vs. M	O vs. K	E vs. A	G vs. C

Table 5-1: Pairwise discriminations used for generalization testing.

Results and Discussion

Acquired Distinctiveness

The first issue explored here was whether or not categorical training led to AD. Evidence for this effect would be provided if perceptual discriminations along a categorization-relevant dimension are better than those of the corresponding controls. In the simulations, for group 1 there are four comparisons that should become sensitized if relevant dimensions acquire distinctiveness: discriminations 1, 2, 3, and 4. For group 2 the four comparisons are 5, 6, 7, and 8. Figure 5-5 shows generalization for each pair, averaged over the four discriminations. The control group showed more generalization between discriminanda for both dimensions, illustrating that for both categorization groups there is evidence for AD between

values of the categorization-relevant dimension that belong to different categories. These results concur with those of Goldstone.

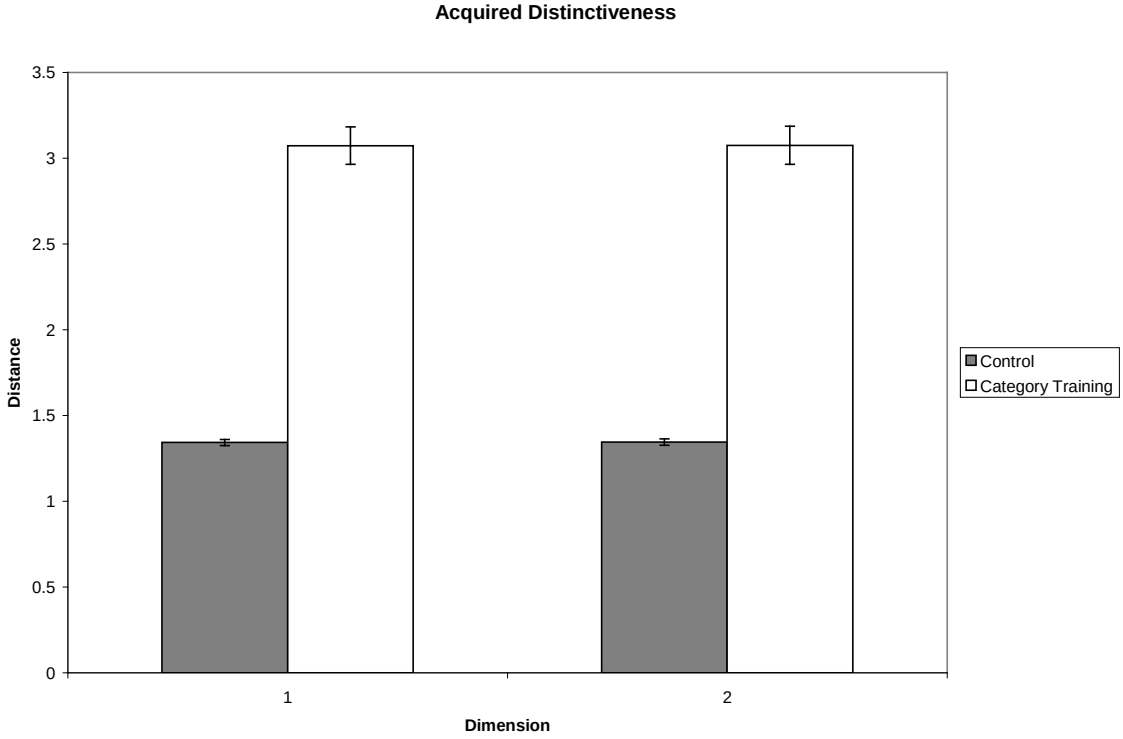


Figure 5-5: Acquired Distinctiveness. Evidence for AD would be provided if perceptual discriminations along a categorization-relevant dimension are better than those of the corresponding controls. In the simulations, for group 1 there are four comparisons that should become sensitized if relevant dimensions acquire distinctiveness: discriminations 1, 2, 3, and 4. For group 2 the four comparisons are 5, 6, 7, and 8. This figure shows the activation of the winning units for one stimulus of the pair when presented with the second stimulus, averaged over the four discriminations and the two stimuli. The control group showed more generalization between discriminanda for both dimensions, illustrating that for both categorization groups there is evidence for AD between values of the categorization-relevant dimension that belong to different categories. These results concur with those of Goldstone.

Acquired Equivalence of the Irrelevant Dimension

Acquired equivalence of a dimension that is irrelevant to the categorization occurs if discriminations along the dimension are worse than those of controls. The simulation data are shown for the two dimension categorization groups in Figure

5-6. The data consist of the average generalization between discriminations 5, 6, 7, and 8 for group 1 and discriminations 1, 2, 3, and 4 for group 2.

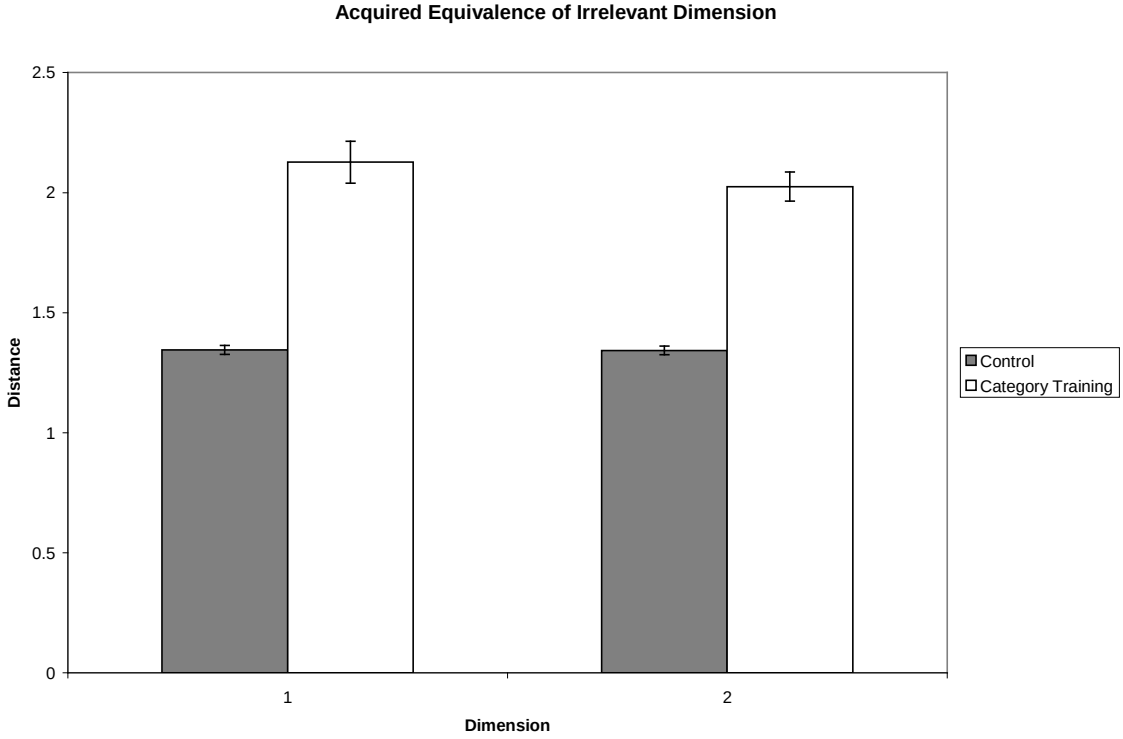


Figure 5-6: Acquired distinctiveness along the irrelevant dimension. The data consist of the average generalization between discriminations 5, 6, 7, and 8 for group 1 and discriminations 1, 2, 3, and 4 for group 2. The degree of acquired distinctiveness is much less substantial than that seen for the relevant dimension.

Instead of demonstrating acquired equivalence, the simulation shows acquired distinctiveness along the irrelevant dimension. The degree of AD is much less substantial, however, than that seen for the relevant dimension. Goldstone found a similar effect when he used stimuli from integral dimensions (saturation and brightness). He found no effect when he used stimuli from separable dimensions (brightness and size).

Acquired Equivalence within the Relevant Dimension

AE within a categorization relevant dimension would occur if stimuli that have different values along a dimension but belong to the same category become less

discriminable with categorization training. For the simulations, this comparison was made across four discriminations for each dimension, each of which involved discriminating between two stimuli that were of the relevant dimension but did not straddle the categorization boundary. In other words, these were stimuli whose values were different but categories were the same. For group 1, I looked at discriminations 9, 10, 11, and 12 and for group 2 I looked at discriminations 13, 14, 15, and 16. Overall, for both dimensions, the degree of generalization for networks in the control condition was greater than the degree of generalization for networks in the categorization conditions (see Figure 5-7). This difference is in the opposite direction as that hypothesized by AE within a dimension, and reproduces the effect obtained by Goldstone.

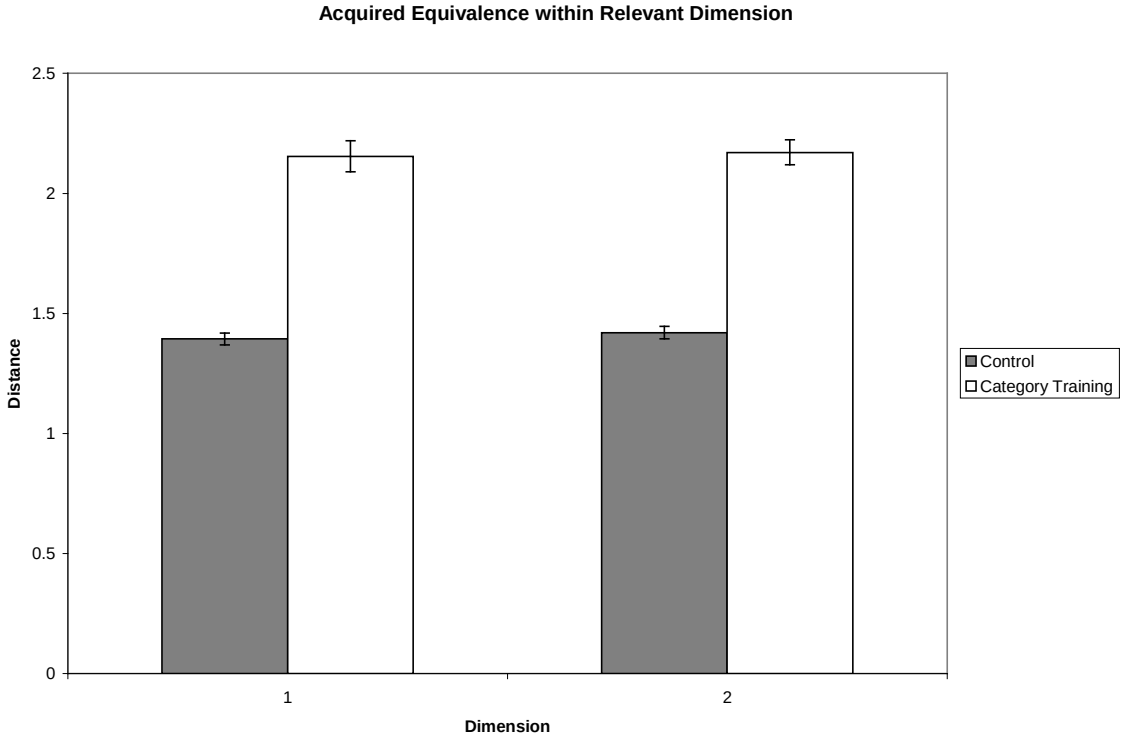


Figure 5-7: Acquired distinctiveness within a relevant dimension. For this simulation, generalization was compared across four discriminations for each dimension, each of which involved discriminating between two stimuli that were of the relevant dimension but did not straddle the categorization boundary. In other words, these were stimuli whose values were different but categories were the same. For group 1, I looked at discriminations 9, 10, 11, and 12 and for group 2 I looked at discriminations 13, 14, 15, and 16. Overall, for both dimensions,

the degree of generalization for networks in the control condition was greater than the degree of generalization for networks in the categorization conditions.

Local Sensitization of a Dimension

Given that acquired distinctiveness was found for category-relevant values whether or not they straddled the category boundary, Goldstone next looked at whether these values were equally sensitized. To test this, discriminations between stimuli that belong to different categories were compared to discriminations between stimuli that belong to the same category but vary along the categorization-relevant dimension. Thus I compared generalization across stimulus pairs that straddled the categories (the "critical" condition) to generalization across stimulus pairs that were within the same category along the relevant dimension (the "noncritical" condition).

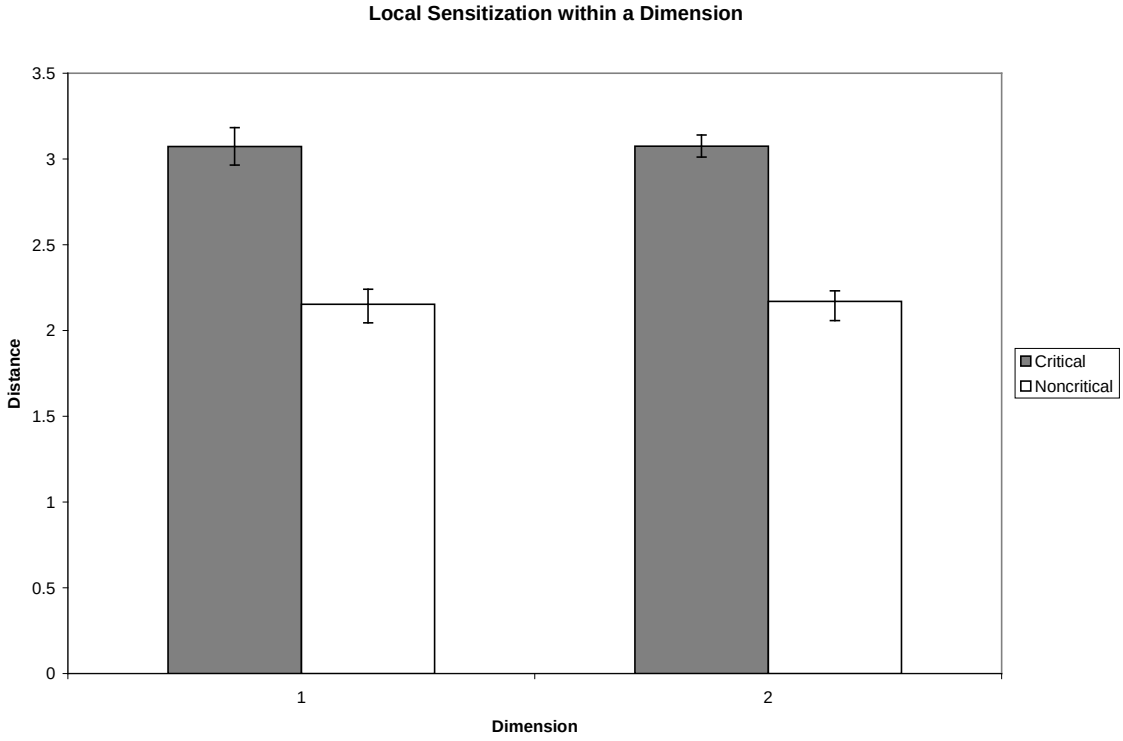


Figure 5-8: Local sensitization within a dimension. Discriminations between stimuli that belong to different categories were compared to discriminations between stimuli that belong to the same category but vary along the categorization-relevant dimension. For group 1, the critical condition consisted of discriminations 1, 2, 3, and 4 and the noncritical condition consisted of discriminations 9, 10, 11, and 12. For group 2, the critical condition consisted of

discriminations 5, 6, 7, and 8 and the noncritical condition consisted of discriminations 13, 14, 15, and 16.

For group 1, the critical condition consisted of discriminations 1, 2, 3, and 4 and the noncritical condition consisted of discriminations 9, 10, 11, and 12. For group 2, the critical condition consisted of discriminations 5, 6, 7, and 8 and the noncritical condition consisted of discriminations 13, 14, 15, and 16. As shown in Figure 5-8, there was more generalization between stimuli in the noncritical condition than between stimuli in the critical condition. Thus, the extended model produces local sensitization within a dimension: although the entire dimension appeared to acquire distinctiveness, the stimuli that straddled the categorization boundary acquired more distinctiveness than those that did not.

General Discussion

The simulations presented in Experiment 1 and Experiment 2 demonstrate that the basic effects of acquired distinctiveness can be accounted for simply by extending the basic differentiation model such that a representation of the reinforcer (or category label) is combined with the representation of the stimulus. Furthermore, the network demonstrated the same types of perceptual sensitization after experience in the acquisition of new categories as did Goldstone's subjects; training networks on different categorization rules resulted in different abilities to make perceptual discriminations

One interesting aspect of Goldstone's results is that he found consistent support for AD, even in cases in which he had expected AE to be produced. Secondly, whereas Goldstone did find that sensitization occurred across an entire dimension, the degree of sensitization depended on the proximity of the stimulus to the categorization boundary. This provides evidence against the standard model of dimensional attention (e.g., Kruschke, 1992; Sutherland & Mackintosh, 1971) in which equivalent sensitization to the entire dimension is predicted. The fact that the current model can account for these data suggests that there may be no high-level

attentional effects at work to produce acquired distinctiveness. Instead, it may simply be the result of a differentiation process that operates on an event representation rather than a stimulus representation.

Another interesting aspect of Goldstone's results concerns the bit of evidence that he did find for acquired equivalence. He found AE of the irrelevant dimension for only one dimension when the dimensions were separable (size and brightness) and for neither dimension when the dimensions were integral (hue and saturation). Goldstone explains this outcome in terms of dimensional attention by suggesting that if two dimensions are naturally attended simultaneously then requiring that attention be placed on one should not necessarily cause the other dimension to be processed less. If the dimensions cannot be simultaneously attended, then desensitization (acquired equivalence) of competing dimensions would be expected when attention is required elsewhere. Thus he suggests that acquired equivalence occurs as a result of divided attention. As was demonstrated in the simulation, the current model has no mechanism for producing AE. From Goldstone's analysis and the data produced by the current model it seems plausible that acquired distinctiveness and acquired equivalence, instead of being opposite sides of the same coin as they are generally considered, instead they may be very different processes that occur at much different levels. AD may be a low-level differentiation process, whereas AE may be a much higher level attentional resource process.

The architecture of the extended model is interesting because, while accounting for acquired distinctiveness effects, it may also have implications with respect to the relationship between learning and memory. For example, it is well known that reacquisition of a simple conditioning task after extinction occurs more quickly than did the initial acquisition (Konorski & Szwejkowska, 1950; Konorski & Szwejkowska, 1952). The current model would explain this as follows. During the initial acquisition of the CR, a strong event layer representation of the stimulus at $<x,y,1>$ will be formed. During extinction, a second representation at $<x,y,0>$ will develop and the overall CR will decrease as the associative weights of the second

representation develop enough to outweigh the effect of generalization to the initial representation. During reacquisition of the CR, the weights of the <*x,y,1*> representation will not have to redevelop from scratch but will merely have to be sufficient to outweigh generalization from <*x,y,0*>. This may have implications related to retention and recall issues such as spontaneous recovery and occasion setting.

CHAPTER 6 APPLICATION TO NEUROPSYCHOLOGY: EFFECTS OF PERIRHINAL CORTEX LESIONS ON STIMULUS REPRESENTATIONS

Introduction

The perirhinal cortex (PRh)—a polymodal region located in the anterior portion of the temporal lobe—is a site of convergence of distinct brain systems involved in memory and perception. For example, PRh has been considered to be part of a functionally unitary medial temporal lobe memory system, damage to any part of which can result in "declarative" memory impairments (Eichenbaum, Otto, & Cohen, 1994; Squire & Zola-Morgan, 1991; Zola-Morgan, Squire, & Ramus, 1994). An alternative view has been that PRh is important for visual information processing leading to "object identification" (Buckley & Gaffan, 1998b; Murray, Malkova, & Goulet, 1998).[7]

Because different researchers have assigned either perceptual or cognitive functions to PRh, it is a particularly interesting brain region when it comes to trying to understand the neural underpinnings of perceptual learning. The dual functions proposed seem to support the idea that the boundary between perception and cognition is not as clear as it was once assumed. Over the past several years, researchers have been trying to determine whether PRh is best characterized as being

[7] For a more detailed discussion of PRh see Appendix B.

specialized for a specific mnemonic function or whether PRh is more simply involved in visual information processing. Resolving this issue has been difficult, in part due to the diverse and puzzling pattern of effects observed following lesions of PRh. It is well established that PRh has an important role in recognition memory for complex objects (Aggleton, Keen, Warburton, & Bussey, 1997; Bussey, Muir, & Aggleton, in press; Ennaceur, Neave, & Aggleton, 1996; Meunier, Bachevalier, Mishkin, & Murray, 1993; Mumby & Pinel, 1994), thought to be a paradigm example of declarative memory (Squire & Zola-Morgan, 1991). More recent research has shown, however, that the use of object recognition paradigms is not necessary to obtain PRh lesion-induced deficits. Buckley and Gaffan (1997), for example, using a large set of object stimuli in a concurrent discrimination paradigm, found that lesions of PRh disrupted retention of preoperatively learned stimuli as well as the acquisition of new discriminations. These results suggest that a general function of PRh might be the discrimination of visual stimuli. However, visual discrimination learning can be preserved following lesions in this region, provided small object sets are used (e.g., Aggleton et al., 1997; Buckley & Gaffan, 1997; Buckley, Gaffan, & Murray, 1997; Bussey, Duck, Muir, & Aggleton, 1999; Bussey et al., in press; Gaffan & Murray, 1992; Rothblat, Vnek, Gleason, & Kromer, 1993; Thornton, Malkova, & Murray, 1998; Thornton, Rothblat, & Murray, 1997). At the same time Buckley and Gaffan (in press), using a relatively small stimulus set size, found that PRh lesions impaired a "configural" discrimination learning task. Furthermore, a consistent yet puzzling effect of PRh lesions is that retention of pre-operatively learned discriminations can be disrupted while the acquisition of new discriminations is not (Buckley & Gaffan, 1997; Eacott, Gaffan, & Murray, 1994; Gaffan & Murray, 1992; Horel, 1994; Horel & Stegner, 1993; Thornton et al., 1997). These inconsistencies in the effects of lesions in PRh require resolution.

In this chapter I investigated whether such effects following lesions in PRh can be accounted for by assuming PRh to have visual information processing properties similar to those in other regions of IT. I constructed a neural network model based on two general assumptions regarding the anatomical and electrophysiological

properties of IT. First, I assumed that networks of neurons in IT set up and maintain representations of visual stimuli, and that these representations can change as a result of experience (perceptual learning), thus facilitating visual discrimination. Accordingly, the model in this chapter is based on the neural network model presented in Chapter 3. Second, I assumed that visual representations in IT are organized in a hierarchical manner, in which simple representations in early processing regions combine to form more complex representations in later, downstream regions. In this way more complex representations are built up as information progresses through the ventral visual stream, culminating in maximum complexity in PRh (perhaps resulting in gestalt representations of complete visual stimuli). I hypothesized that a model embodying these simple functional and organizational principles should be able to account for effects of lesions in PRh.

In order to test this model, lesions were made in the component of the network corresponding to PRh, and the resulting effects compared with the effects of lesions in PRh in monkeys. I have focused on recently reported effects of lesions in PRh in monkeys that have provided key insights into the functions of this region, and that have led to a reappraisal of the function of PRh in terms of "object identification" (Buckley & Gaffan, 1998b; Murray et al., 1998) .

Functional and Organizational Principles

In the present model, I assume that the function of IT is to prepare and store visual stimulus representations which, in cooperation with downstream brain regions, are subsequently associated with responses and reinforcers. Such regions may include the amygdala, basal ganglia, and prefrontal cortex, all of which receive projections from IT (Iwai & Yukie, 1987; Ungerleider, Gaffan, & Pelak, 1989). I suggest that IT accomplishes this stimulus preprocessing through mechanisms of discrimination and perceptual learning.

Lesions in IT lead to impairments in visual discrimination (Desimone, Albright, Gross, & Bruce, 1984; Desimone & Gross, 1979; Tanaka, 1996). In order to

investigate whether activity of cells in IT is correlated with discrimination behavior, Jagadeesh and Desimone (1997) created a testable dimension by isolating stimuli that were "good" or "poor" for a given cell in PRh and then "morphing" these images together in various proportions to generate a continuum of stimuli ranging between the two extremes. Using a pair-wise discrimination paradigm it was found that neurons' discrimination learning curves for these stimuli matched the behavioral discriminations almost identically. This result suggests that stimulus-selective neural activity in IT drives behavioral selection of visual stimuli in discrimination learning (see also Kobotake, Wang, & Tanaka, in press).

Perceptual learning can be thought of as the changes in stimulus representations that occur with experience. Thus the development of stimulus representations as manifested in the selective firing of neurons is an example of perceptual learning, and the existence of neurons that have learned to fire to particular stimuli is evidence that perceptual learning has occurred. These learned responses, however, must be distinguished from innate, hard-wired patterns of neuronal selectivity. Evidence that experience may be sufficient to alter representations in IT has been obtained in several studies (e.g., Booth & Rolls, 1998; Logothetis & Pauls, 1995; Sakai & Miyashita, 1991; Tanaka, 1993). Furthermore, Tovee, Rolls, and Ramachandran (1996) demonstrated perceptual learning by neurons in IT. These authors first located a population of face-sensitive cells. The monkeys were then presented with an ambiguous (degraded) picture of a face to which the neuron did not respond. The unambiguous version of that same face was then presented, and the neuron responded. Finally, in the critical test the same ambiguous face was re-presented. The neuron now responded to this previously ambiguous stimulus. This suggests that the neuron learned to identify a degraded form of the stimulus following preexposure to a complete version.

The foregoing studies suggest that discrimination and perceptual learning are processes intrinsic to IT. The present model thus takes as its starting point the basic competitive network, discussed earlier in this dissertation, that performs these

functions in a manner consistent with extant data on discrimination and perceptual learning in animals.

In the present model the visual preprocessing in IT is viewed as nonassociative: No reinforcement signal is required to instruct representational changes. Instead, this occurs in an automatic, "unsupervised" manner. This assumption seems reasonable given that changes in visual representations—as evidenced by changes in selectivity of neuronal encoding of visual stimuli in IT—can occur during passive viewing of visual stimuli, without reinforcement and outside of the context of a "task" (Brown, 1996). Furthermore, certain changes in neuronal responding to visual stimuli have been observed even in anaesthetized animals (Miller, Li, & Desimone, 1991).

As a visual preprocessor, an intact IT will be necessary for the performance of many visual tasks, such as discrimination learning and categorization. However, while lesions within IT often produce deficits in such tasks, equally often they do not. I propose that this can be explained by considering the hierarchical organization of the ventral visual stream, in which neurons code representations of increasingly complex stimuli (Desimone & Ungerleider, 1989). This complexity appears to reach its maximum in PRh, and there is evidence that this complexity can involve the "fusion" of visual stimuli (Tanaka, 1996). Thus while lesions in PRh cortex will abolish representations of certain complex stimuli, representations of the features that comprise those stimuli will be preserved in upstream regions of IT such as TE. One consequence of this is that stimuli that differ according to these simple features will be discriminable even without the complex representation stored in PRh. The implications of this will become clearer when I consider how the model accounts for some known effects of lesions in PRh.

In the following section I describe the architecture of the model, which is based on the differentiation network that is described in Chapter 3. The differentiation network is henceforth referred to as the "basic component network".

Architecture of the Model

As outlined above, the model is based on several assumptions, including (1) IT supports discrimination and perceptual learning, (2) stimulus representations are built up in a hierarchical manner, and (3) this learning occurs in a nonassociative, unsupervised manner. The first and third features of the model are intrinsic to the basic component network. The second is a consequence of how the basic component networks are arranged within the full architecture of the model (see Figure 6-1). Both the Feature Conjunction Layer and the Feature Layers have the properties of the competitive layer in the basic component network, with the exception that the Feature Conjunction Layer receives input from two Feature Layers (F_1 and F_2), which receive input from two input layers (I_1 and I_2). F_1 is fully connected to I_1 and F_2 is fully connected to I_2. The Input and Feature Layers consist of seven-by-seven unit grids. The Feature Conjunction Layer consists of a four-dimensional space of units (seven-by-seven-by-seven-by-seven). All three of these layers are fully connected with the outcome representation. As in the basic component network, the nonassociative preprocessor represents IT and the associative processor represents downstream brain areas involved in associating visual stimulus representations in IT with reinforcers and responses. *The Feature Conjunction Layer represents PRh and the Feature Layers represent upstream areas in IT such as area TE.*

In the full architecture, a stimulus comprised of two features can be fed into the network. Stimuli consist of vectors of quadruples of values (e.g., *<x1,y1,x2,y2>*). There are two values within the quadruple for each of the two features. Thus if a stimulus consisting of a large red rectangle is presented to the network, then *<x1,y1>* might represent hue and saturation of the color and *<x2, y2>* might represent length and width of the rectangle. By way of the mechanisms detailed in the previous section, these features come to be represented in the two Feature Layers. The Feature Layers converge on the Feature Conjunction Layer such that a complex stimulus consisting of the two features is represented. Note that both of the features

of this complex stimulus in the Feature Layer as well as the full representation in the Feature Conjunction Layer can be associated with an outcome (reinforcer or response) via direct, full connectivity with the output node.

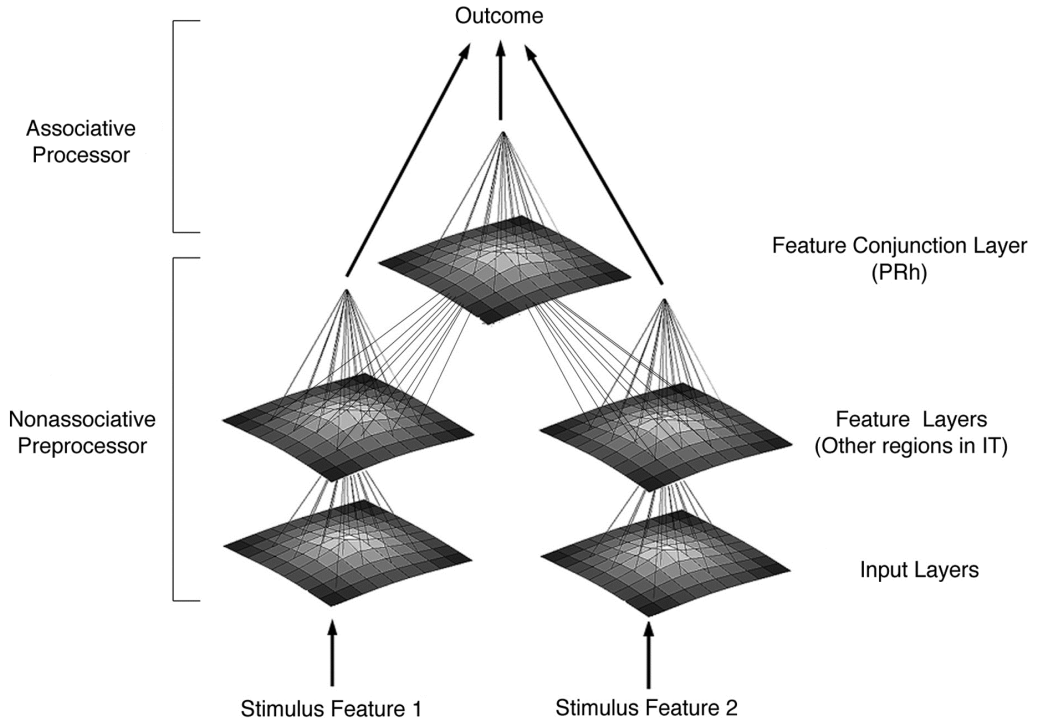


Figure 6-1: The full architecture of the model. Two separate input layers (I_1 and I_2) are utilized. Each input layer represents one feature of a complex stimulus. When a stimulus is input to the network, each of the input layers is activated depending on the value of the stimulus on the particular feature that it represents. The competitive portion of this network consists of 3 separate groups of units. The Feature Layers (F_1 and F_2) consist of two groups of units, one associated with feature 1 and one associated with feature 2. F_1 is fully connected to I_1 and F_2 is fully connected to I_2. The Feature Conjunction Layer is fully connected with both F_1 and F_2, and therefore can represent the conjunction of the features represented in these layers. All three of these layers are fully connected with the outcome representation. As in the basic component network, the nonassociative preprocessor represents IT and the associative processor represents downstream brain areas responsible for associating visual stimulus representations in IT with reinforcers and responses. Within the nonassociative preprocessor, the Feature Conjunction Layer represents PRh and the Feature Layers represent upstream areas in IT such as area TE.

General Methods

The experiments in the current study compare the performance of intact networks to those that have been "lesioned". In the present model, the Feature Conjunction Layer occupies the same relative position as PRh in the ventral visual processing stream. Thus, removal from the network of the Feature Conjunction Layer corresponds to a lesion of PRh. When the network is lesioned, the Feature Layers are still fully functional since neither their inputs from the Input Layers nor their outputs to the Outcome representation are affected by the lesion.

Identical network parameters were used for every simulation presented in this paper. In each experiment, five simulations were run in both the lesioned and intact conditions. In this way the experiments are designed in a manner similar to lesion experiments in animals, comparing the performance of a "Lesion" group (n=5) with that of an intact "Control" group (n=5). The specific parameters used were as follows: λ = 1.0 or 0.0,δ = 0.05, α = 0.05, β= 0.005, σ= 1.0.

Experiment 1: Stimulus Set Size Effects in Concurrent Discrimination Learning

Much evidence points to a role for PRh in the learning and retention of visual discriminations. Lesions of PRh have been shown significantly to disrupt learning and retention of visual pair-wise discriminations when a large number of stimulus pairs must be discriminated concurrently (Buckley & Gaffan, 1997). PRh does not, however, appear to have a general role in discrimination learning. Several studies using small stimulus set sizes, in monkeys as well as in rats, have found preserved discrimination learning following lesions that include this region (e.g., Aggleton et al., 1997; Buckley & Gaffan, 1997; Buckley et al., 1997; Bussey et al., 1999; Bussey et al., in press; Gaffan & Murray, 1992; Rothblat et al., 1993; Thornton et al., 1998; Thornton et al., 1997). Similar set-size effects are observed in the delayed nonmatching-to-sample paradigm (Eacott et al., 1994).

In this experiment I investigated the effects of stimulus set size in a concurrent learning paradigm similar to that used by Buckley and Gaffan (1997).

Methods

Three sets of stimulus pairs were created by randomly selecting a quadruple of numbers for each stimulus. Set "Small" consisted of one pair of stimuli, set "Medium" consisted of five pairs of stimuli, and set "Large" consisted of ten pairs of stimuli. Two groups of five networks each were then initialized: Group "Control" consisted of intact networks whereas group "Lesion" consisted of networks that had been lesioned in order to simulate the effect of a lesion in PRh. Each network was trained on a pairwise concurrent discrimination task using set "Small" until it reached criterion (90% correct on five consecutive sessions). The networks were reinitialized to prevent interference effects across stimulus sets, after which each network was trained to criterion on the five pair concurrent discrimination (set "Medium"). The networks were again reinitialized, and then were trained to criterion on the ten pair concurrent discrimination (set "Large").

Results and Discussion

As shown in Figure 6-2, group Lesion was impaired relative to group Control on concurrent discrimination learning when the stimulus set size was medium or large, but not when it was small. Two-way analysis of variance with Group and Set Size as factors revealed a significant main effect of Group, $F(1,8)=50.9, p=0.0001$, a significant main effect of Set Size, $F(2,16)=106.3, p<0.0001$, and a significant Group x Set Size interaction, $F(2,16)=41.1, p<0.0001$. Analysis of simple main effects revealed that group Lesion committed significantly more errors in reaching criterion than did group Control in the large set size condition, $F(1,8)=46.3, p<0.001$, and in the medium set size condition, $F(1,8)=11.6, p<0.01$, but not in the small set size condition, $F(1,8)=1.31, p=0.29$. Both groups committed more errors as the set size increased, as indicated by a significant effect of Set Size for both group Control, $F(2,16)=8.02, p<0.01$, and group Lesion, $F(2,16)=139.4, p<0.001$.

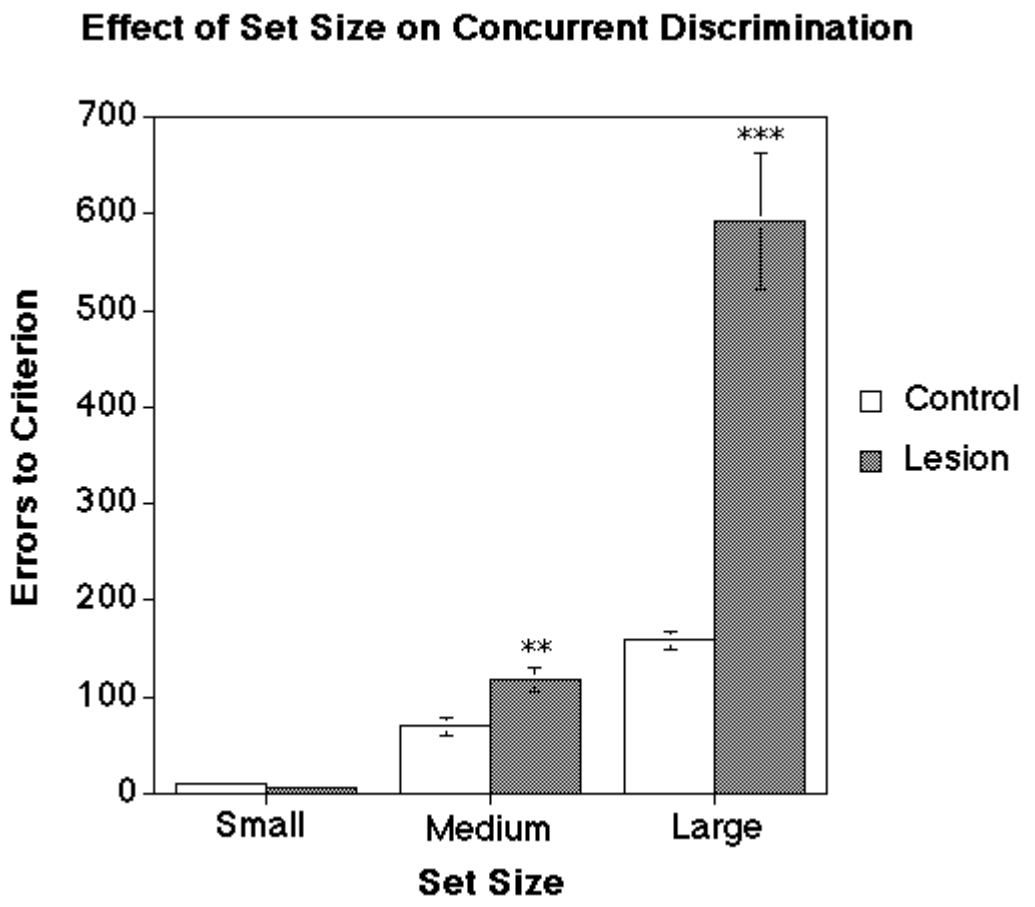


Figure 6-2: Effect of set size on acquisition of pairwise concurrent discrimination. Error bars represent +/- SEM. Asterisks indicate significant difference from group Control; ** p<0.01; *** p<0.001.

This experiment shows that lesioning the Feature Conjunction Layer of the network can reproduce stimulus set-size effects similar to those observed following lesions of PRh in monkeys (Buckley & Gaffan, 1997). Specifically, the larger the set size, the greater the effect of the lesion. This occurs because whereas the intact networks can represent individual features of a stimulus as well as their conjunction (e.g., they can represent "red", "square", and "red & square"), the lesioned networks can represent only the individual stimulus features ("red" and "square"). In this sense, the current model of PRh is similar to Gaffan, Harrison, and Gaffan's (1986) computational model of IT, in which lesions in IT are thought to reduce the number of "stimulus attributes" available during visual discrimination learning. As a result,

due to their richer stimulus representation the intact networks can build several associative links (red→+; square→+; red&square→+) whereas the lesioned networks can only build associative links from the individual feature representations to the outcome (red→+; square→+). As the stimulus set size increases, the probability increases that a particular feature will be part of both a reinforced and a non-reinforced stimulus. The probability of a specific conjunction of features being both reinforced and non-reinforced, however, is much smaller. Thus an intact network, and an animal with an intact PRh, will have an advantage over their lesioned counterparts because they can represent separately the conjunction of stimulus features. In contrast the lesioned network, or animal, must rely on the spared individual stimulus features to attempt to solve the discrimination. This advantage increases as the degree of feature overlap increases. Since the probability of feature overlap increases with set size, so does the advantage of the intact group.

In this sense large-set concurrent discrimination is similar to "configural" tasks in which overlap of features is at a maximum. To illustrate, if I represent features by the letters A,B,C, and D, objects by the conjunction of two of these features (e.g., AB), reinforcement by +, and nonreinforcement by -, then an example of such a configural task is the "biconditional discrimination task": AB+, AC-, CD+, AD-. In discriminating a rewarded stimulus conjunction from a nonrewarded one, the probability of feature overlap is one. Therefore, it would be predicted that, given stimuli of appropriate complexity, PRh lesions would have a severely disruptive effect on the acquisition of such a task. This issue is addressed in Experiment 2.

A central theme of the present study is that it may not be necessary to assign a special function for PRh distinct from that of other areas of IT—in specific mnemonic or associative processes, for example—to account for effects of PRh lesions on tasks such as concurrent discrimination. Instead, whether or not deficits are observed following a lesion in PRh will depend on the stimulus demands of the task. Furthermore, provided an appropriate level of stimulus complexity, lesions in IT that spare PRh should also disrupt concurrent discrimination learning. Indeed, just such a result has been reported (Iwai & Mishkin, 1968; Malkova, Mishkin, &

Bachevalier, 1995; Moss, Mahut, & Zola-Morgan, 1981; Phillips, Malamut, Bachevalier, & Mishkin, 1988).

It was once thought that the critical factor in explaining lesion-induced deficits in concurrent learning was not the stimulus demands of the task, but the intertrial interval. Specifically, a version of the concurrent task that featured an intertrial interval of 24 hours was viewed as a "habit" (or "procedural") task that should be spared by a lesion of PRh and other components of the "medial temporal lobe memory system", but may be impaired by lesions in other regions of IT (Phillips et al., 1988). Buckley and Gaffan (1997), however, obtained their PRh lesion-induced deficits using a 24-hour version of the concurrent discrimination task. Thus it appears that, consistent with the present model, the idea that PRh can be distinguished from other regions in IT according to a memory/habit (or declarative/procedural) dichotomy is in need of revision (Buckley & Gaffan, 1997).

Experiment 2: Configural Learning

In Experiment 1 it was shown that that lesioning the Feature Conjunction Layer of the network can reproduce stimulus set-size effects similar to those observed following lesions of PRh in monkeys. According to the present model this is because the representations of feature conjunctions which are normally used to represent and discriminate objects are unavailable to the animal (or network), and it must rely on the spared simple feature representations to attempt to solve the discrimination. In situations such as large-set concurrent learning, stimuli are more likely to have features in common, and thus the lesioned animal or network will experience much greater inter-item interference.

Such feature overlap reaches a maximum in "configural" tasks in which items cannot be discriminated according to simple features, but rather can only be discriminated through the use of representations of the conjunction of features. An example of such a task is the biconditional discrimination task. The biconditional discrimination task can be formalized as AB+, AC-, CD+, AD- (where features are

represented by the letters A,B,C, and D, objects by the conjunction of two of these feature (e.g., AB), reinforcement by +, and nonreinforcement by -). It can be seen that this task cannot be solved by assigning reinforcement or nonreinforcement to any of the features A,B,C or D, as each of these features is associated equally with both reinforcement and nonreinforcement. Only by associating the conjunctions of features (e.g., AB) with reinforcement can the problem be solved. For this reason it is predicted that lesioning the Feature Conjunction layer of the network should have a devastating effect on the acquisition of such tasks. Similarly, given stimuli of appropriate complexity, lesions in PRh in monkeys should lead to deficits in configural learning. Furthermore, because of the high degree of overlap of features in such tasks, deficits should be obtainable using a smaller set size than is necessary to produce deficits on a concurrent learning task. Indeed precisely this finding has recently been reported by Buckley and Gaffan (in press) using an extended version of the biconditional discrimination task. The present experiment tests these predictions using the standard version of the task, AB+, AC-, CD+, AD-.

Methods

Four stimulus features were created: A=<1,7>; B=<7,7>; C=<1,1>; D=<7,1>. These parameters were chosen in order to maximize the discriminability of the features A, B, C, and D by maximizing the Euclidean distance between them on the input grid. These features were then combined to produce a biconditional discrimination as follows:

Pair 1: AB (<1,7> <7,7>) → + AC (<1,7> <1,1>) → -

Pair 2: CD (<1,1> <7,1>) → + AD (<1,7> <7,1>) → -

Two groups of 5 networks each were then initialized: Group "Control" consisted of intact networks whereas group "Lesion" consisted of networks that had been lesioned in order to simulate the effect of a PRh lesion. Each network was trained on the biconditional pairwise concurrent discrimination task until it either

reached criterion (90% correct on five consecutive sessions) or exceeded 400 sessions, at which point it was deemed to have "failed" the discrimination.

Results and Discussion

Lesioning the Feature Conjunction Layer of the network had a devastating effect on the acquisition of the biconditional discrimination (see Figure 6-3). Analysis of variance revealed a highly significant effect of Group, $F(1,8)=297.4, p<0.0001$. All networks in group Lesion failed, taking more than 400 sessions to attain criterion.

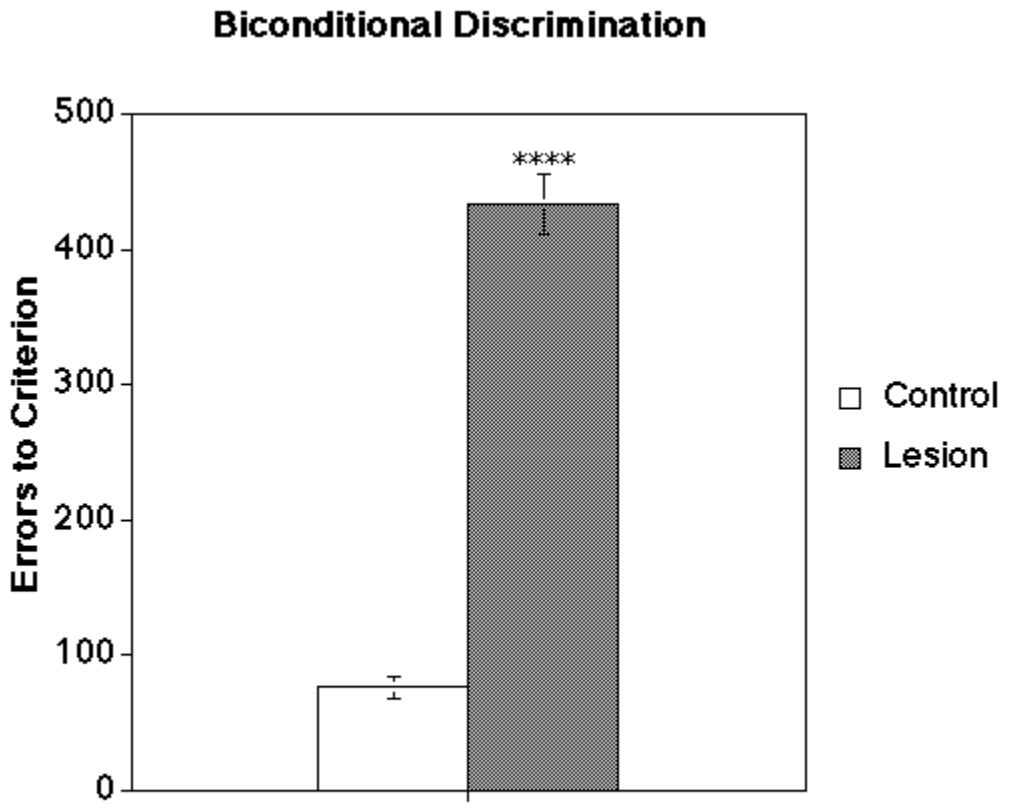


Figure 6-3: Acquisition of a biconditional discrimination. Error bars represent +/- SEM. Asterisks indicate significant difference from group Control; **** p<0.0001.

Lesioned networks were severely impaired on the biconditional discrimination. Although the biconditional discrimination set size was considerably smaller than that used for the large set concurrent discrimination (Experiment 1), the lesioned networks showed a relatively greater impairment in acquisition of the biconditional discrimination task. This is because the biconditional discrimination task shifts the probability of feature overlap to one: Each feature is both reinforced and nonreinforced, thereby eliminating the need for a large set size in order to bring out a deficit in the lesioned networks. Concurrent discriminations are, however, very difficult for the lesioned network and for PRh-lesioned animals when the probability of feature overlap is high (i.e., when the stimulus set-size is large; Experiment 1; Buckley & Gaffan, 1997). In this sense a large set-size concurrent discrimination can be thought of as a "partial" configural task.

In the present model, PRh is represented by the Feature Conjunction Layer of the network, which contains configural representations of stimuli constructed from features in upstream areas in IT. This allows the network to simulate lesion effects on the biconditional discrimination task; and on concurrent discrimination with a large set size (Experiment 1). However, it is important to emphasize that I do not view PRh as specialized for configural learning as opposed to "elemental" learning, in a manner akin to that proposed for the hippocampus (Sutherland & Rudy, 1989). In the present model a particular region of IT is configural only *relative* to areas immediately upstream, which could be thought of as containing elemental representations. That elemental area, however, would serve as a configural layer for the area immediately upstream from it, and so no region of IT can be considered exclusively configural or elemental, and deficits on "configural" tasks will depend on the complexity of the stimulus material. (Consistent with this, Bussey, 1997, using simple light and tone stimuli, found no effect of lesions including perirhinal cortex on configural learning in the rat.) Thus it may not be necessary to propose distinct configural versus elemental learning systems in the brain to account for deficits on "configural" tasks (Bussey, Warburton, Aggleton, & Muir, 1998). For the same reasons, the present view is not strictly aligned with either configural (Pearce, 1987b;

Pearce, 1994) or elemental (Rescorla & Wagner, 1972) psychological theories of learning. It is, however, close in spirit to Pearce's view that all stimulus representations should be regarded as configural.

Experiment 3: Greater Effects of Lesions in PRh on Retention versus Acquisition

A consistent yet puzzling effect of PRh lesions is that retention of pre-operatively learned discriminations can be disrupted while the acquisition of new discriminations is not (Buckley & Gaffan, 1997; Eacott et al., 1994; Gaffan & Murray, 1992; Horel, 1994; Horel & Stegner, 1993; Thornton et al., 1997). Only when large stimulus set sizes are used are concurrent discriminations impaired (Buckley & Gaffan, 1997). Somehow reacquisition and new learning are differentially affected by PRh lesions (Buckley & Gaffan, 1997). Some authors have attempted to explain this phenomenon by appealing to separate object-knowledge (PRh) and procedural (non-PRh) systems in the brain (Thornton et al., 1997). I predicted that that the simple properties intrinsic to the present model could account for this result.

In this experiment I test the prediction that lesions of the Feature Conjunction Layer of the network should yield greater effects on retention of concurrent discrimination problems than on acquisition.

Methods

Four sets of stimuli were created, two consisting of two randomly selected pairs of stimuli each and two consisting of ten randomly selected pairs of stimuli each. Two groups of five networks each were then initialized. For the first phase of this experiment, both group "Control" and group "Lesion" consisted of intact networks. Each network was trained on a pairwise discrimination task using the first small stimulus set until it reached criterion (90% correct on five consecutive sessions). For the second phase of the experiment, the Feature Conjunction Layer was removed from the networks in the "Lesion" group, and then both groups (using the same

non-reinitialized networks) were trained again to criterion on the first stimulus set. Finally, for the third phase of the experiment both groups (again not reinitialized) were run on a new pairwise discrimination using the stimuli from the second small set. The above procedure was then repeated using the two large stimulus sets.

Results and Discussion

As shown in Figure 6-4, group Lesion was impaired relative to group Control on the retention, but not the acquisition, of a concurrent discrimination when the stimulus set size was small. Two-way analysis of variance with Group and Stimulus Set as factors revealed no significant overall main effect of Group, $F(1,8)=0.89, p=0.37$, a significant main effect of Stimulus Set, $F(2,16)=4.78, p<0.05$, and a significant Group x Stimulus Set interaction, $F(2,16)=4.57, p<0.05$. Analysis of simple main effects revealed that group Lesion committed significantly more errors than did group Control in reattaining criterion on the discrimination learned prior to lesioning, $F(1,8)=10.9, p=0.01$, but the groups did not differ in the acquisition of a discrimination with a new stimulus set, $F(1,8)=0.38, p=0.56$. There was no difference between the groups in the acquisition of the discrimination learned prior to lesioning, $F(1,8)=0.74, p=0.41$.

As shown in Figure 6-5, group Lesion was impaired relative to group Control on the retention *and* acquisition of a concurrent discrimination when the stimulus set size was large. Two-way analysis of variance with Group and Stimulus Set as factors revealed a significant main effect of Group, $F(1,8)=41.1, p<0.001$, no significant main effect of Stimulus Set, $F(2,16)=0.31, p=0.74$, and a significant Group x Stimulus Set interaction, $F(2,16)=13.5, p<0.001$. Analysis of simple main effects revealed that group Lesion committed significantly more errors than did group Control in reattaining criterion on the discrimination learned prior to lesioning, $F(1,8)=25.6, p=0.001$, and in the acquisition of a discrimination with a new stimulus set, $F(1,8)=19.4, p=0.002$. There was no difference between the groups in the acquisition of the discrimination learned prior to lesioning, $F(1,8)=0.18, p=0.68$.

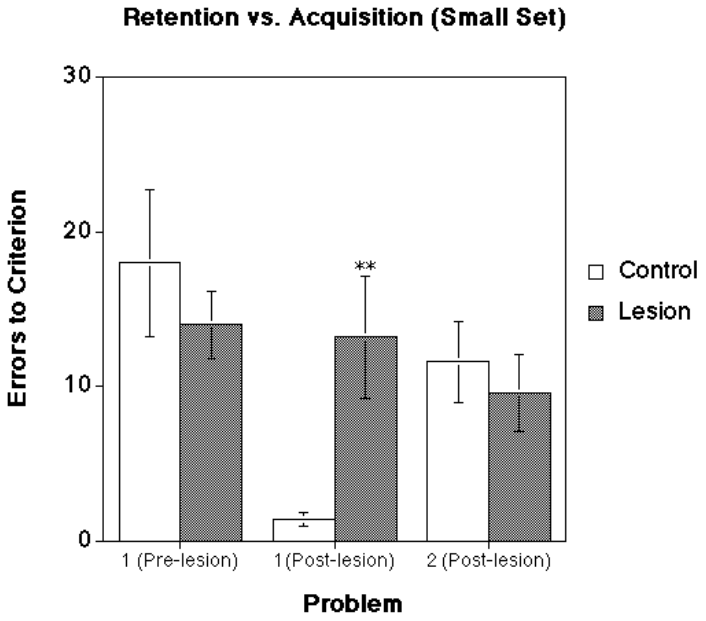


Figure 6-4: Effect of lesioning the Feature Conjunction Layer on retention and acquisition of a small set size pairwise concurrent discrimination. Error bars represent +/- SEM. Asterisks indicate significant difference from group Control; ** p<0.01; *** p<0.001

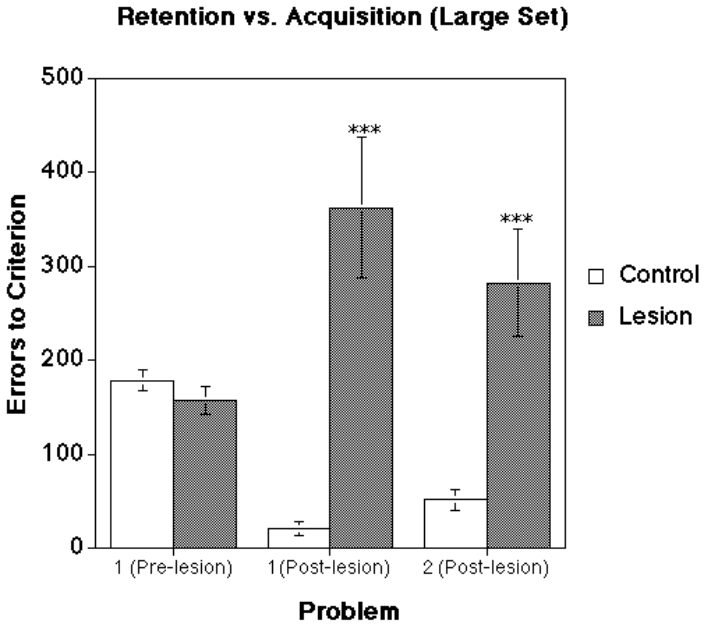


Figure 6-5: Effect of lesioning the Feature Conjunction Layer on retention and acquisition of a large set size pairwise concurrent discrimination. Error bars represent +/- SEM. Asterisks indicate significant difference from group Control; ** p<0.01; *** p<0.001

The present experiment reproduces the effect that following PRh lesions in monkeys, retention deficits are obtained more readily than acquisition deficits. Using a small stimulus set size, the lesioned networks were impaired in the retention of a pre-operatively learned discrimination, but were unimpaired in acquiring a new problem. However, when a large set size was used, the lesioned networks were impaired in both retention of a pre-operatively learned discrimination and acquisition of a new concurrent discrimination. Thus lesions in the network are able to produce deficits in both retention and acquisition, but the former occurs more readily, requiring only a small set size, whereas larger set sizes are needed to produce deficits in acquisition.

The network produces the above pattern of results as follows. During acquisition of a discrimination in an intact network, three representations of the stimulus, or portions of the stimulus, will have been formed: F_1, F_2, and a representation of the complete stimulus in the Feature Conjunction Layer. As a result, associative links will be constructed between the outcome node and each of the three stimulus representations. During training, the associative learning rule operates such that the sum of the weights of the associative links, modulated by the degree of activation of the corresponding stimulus representation units in F_1, F_2, and the Feature Conjunction Layer, asymptotes at a value of one. This value is output as the response of the network. After the network has been trained and then lesioned, however, the stimulus conjunction representation in the Feature Conjunction Layer will no longer exist, and the associative links from the Feature Conjunction Layer will no longer contribute to the response of the network. As a result, the performance of the network on the previously learned discriminations will be impaired.

The reason for this effect in PRh-lesioned monkeys has been puzzling. The present model, however, provides a parsimonious explanation for this effect. Under normal circumstances an intact animal can solve a simple, small set size discrimination using either the representation of the features of the stimulus or the conjunction of those features. Normally the animal will use both: The learning is

distributed across these two types of representation. Thus when PRh is lesioned following acquisition of the discrimination, a major contributor to the associative connections leading to the response is removed. The result is an impairment in the performance of the discrimination relative to an intact animal. If, however, the lesioned animal is then presented with a new simple small set size discrimination, all of the learning can be accomplished using the feature representation alone and there will be no difference between the lesioned and the intact animal in their ability to solve the discrimination (see also Experiment 1).

Experiment 4: Memory Capacity versus Visual Information Processing Accounts of Lesion-Induced Deficits

The finding that lesions in PRh lead to discrimination learning deficits using large sets of stimuli but not small ones (see Experiment 1), as well as the observation of greater effects of PRh lesions of retention versus acquisition (see Experiment 3), suggests a simple explanation: The lesions may reduce the storage capacity of the damaged system. Thus small numbers of stimulus representations can be formed and stored, but large numbers of stimuli would exceed this reduced storage capacity. An experiment by Horel and Stegner (1993), however, suggests that this is not the explanation. These authors trained animals on eight concurrent pair-wise object discriminations and then tested for retention. Monkeys with cooling-induced suppression in IT were impaired in retention, but only on a subset of the discriminations. Taken alone these data support the notion of impaired memory capacity. However Horel and Stegner went on to test these monkeys on the subset of discriminations they had previously failed, with IT both suppressed and fully functional. Cooling again impaired discrimination on the pairs in this subset. If the original deficit had been due to the number of stimulus pairs exceeding the memory capacity of the damaged system, then the animals should have been unimpaired when tested on this smaller subset of problems. This result thus does not support

the idea that the lesion compromised memory capacity *per se*, but instead suggests that the retention deficit was related to the properties of the particular stimuli with which the lesioned animals had difficulty.

It should be noted that it is unclear whether the cooling in Horel and Stegner (1993) affected functioning of PRh. However, this experiment addresses an issue of particular relevance to the present study, namely whether discrimination impairments following lesions in IT, and thus PRh, are better characterized as failures of memory *per se* or of visual information processing. The present model is consistent with the latter view, and thus makes the prediction that, using stimuli of appropriate complexity, a pattern of results similar to that obtained by Horel and Stegner (1993) should be obtained following selective lesions of PRh.

The present model predicts that animals with a nonfunctional PRh should have particular difficulty with discriminations in which there is a high degree of overlap between features. This was illustrated in Experiments 1 and 3, in which lesioned networks were impaired on large set-size concurrent discriminations, and in Experiment 2 in which lesioned networks were impaired on a configural discrimination. This suggests that the particular stimulus pairs on which the lesioned animals in the study of Horel and Stegner (1993) were impaired were those in which the stimuli had a high degree of feature overlap. In this experiment, I test this idea in the context of the experimental design of Horel and Stegner (1993).

Methods

One set of eight stimulus pairs was created by selecting a quadruple of numbers for each stimulus. Each of five networks was then initialized and trained on a pairwise concurrent discrimination using the eight pairs until criterion was reached (seven out of eight correct on five consecutive sessions). Next, the networks were damaged in order to simulate the effect of PRh lesions. The now-lesioned networks were trained again on the same pairwise concurrent discrimination task for a total of 50 trials and the performance of the networks on each stimulus pair was recorded.

The lesions were then reversed and each intact network was tested on only the stimulus pairs on which the lesioned networks had performed at a level significantly below chance. Finally, the five networks were again lesioned and tested on only the stimulus pairs on which they had performed at a level significantly below chance.

Results and Discussion

All intact networks reached the criterion of 87.5% (seven out of eight discriminations) correct on five consecutive sessions on each of the eight problems in a mean of 60.4 sessions. Following lesioning of the Feature Conjunction Layer, performance on three of the problems dropped to below chance levels as indicated by analysis of 95% confidence limits (lower 95% confidence limits: problem three = 27.6%; problem five = 37.8%; problem eight = 27.1%; see Figure 6-6). These were the three most difficult problems for all five networks. The lesion was then reversed and the performance levels of these three problems increased to 84%, 87.2% and 87.2%, respectively (Figure 6-7). When the network was again lesioned, performance dropped to 50.4%, 62.4% and 58.4% (Figure 6-7). Analysis of 95% confidence limits revealed that the performance of the lesioned network was not significantly above chance on problems three and eight (lower 95% confidence limits = 41.5% and 45.4%, respectively) and only marginally above chance on problem five (lower 95% confidence limit = 57.0%). Two-way analysis of variance performed on the data from group Lesion and group Control on problems three, five and eight revealed a significant main effect of Group, $F(1,8)=93.6, p<0.0001$, a main effect of Problem, $F(1,8)=5.67, p<0.05$, and no Group x Problem interaction, $F(1,8)=1.77, p=0.20$. Analysis of simple main effects revealed that group Lesion performed at a significantly lower level than group Control on problem three, $F(1,8)=89.9, p<0.001$, problem five, $F(1,8)=109.8, p<0.001$, and problem eight, $F(1,8)=26.9, p<0.001$.

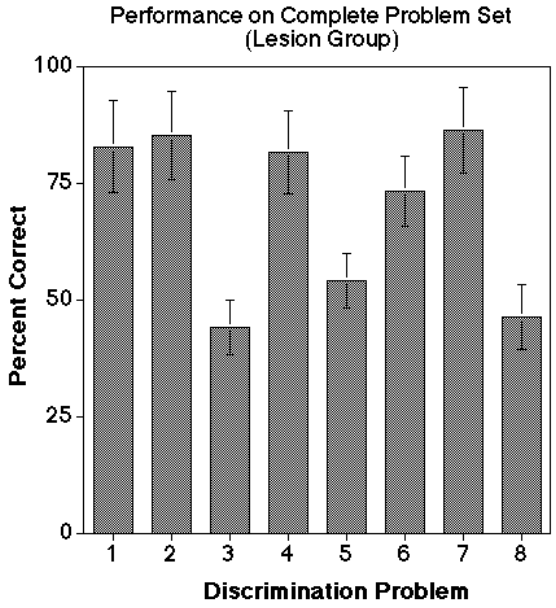


Figure 6-6: Performance of the network on the complete set of eight discrimination problems following the lesion of the Feature Conjunction Layer. Following the lesion, the network failed to perform above chance level (50% correct) on problems 3, 5 and 8.

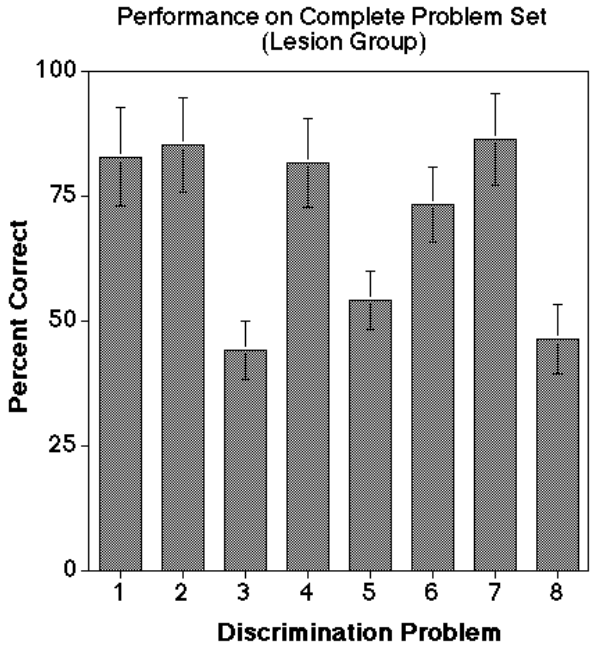


Figure 6-7: Performance of group Lesion and group Control on the three problems (3, 5 and 8) which the lesioned network failed during the original eight-pair concurrent discrimination. Following the lesion, the network again failed to perform above chance level (50% correct) on these problems. Error bars represent +/- SEM. Asterisks indicate significant difference from group Control; *** $p<0.001$.

The results of this experiment explain the effect found by Horel and Stegner (1993) by assuming that certain stimulus pairs which have a high degree of feature overlap will be difficult for the lesioned networks to discriminate. Of the stimuli generated for this simulation, pairs three, five and eight had a particularly high degree of overlap in that seven of twelve "features", either within or across pairs, were identical yet had opposing consequences (stimulus pair 3 = <4 7 2 7> +, <4 5 2 7> -; stimulus pair 5 = <3 6 5 6> +, <4 7 5 6> -; stimulus pair 8 = <4 5 7 1> +, <5 6 7 1> -).

The present experiment thus supports an interpretation of PRh lesion-induced deficits in terms of visual information processing, rather than memory *per se*. It also suggests that PRh lesion-induced deficits on concurrent discriminations (Experiments 1 and 3) may be due to animals failing to discriminate specific pairs of stimuli (I would also expect a perhaps lesser contribution to the deficit from inter-pair interference). Indeed this suggests that it is not the concurrent aspect of such multiple-pair discriminations that is the important factor; the same pairs of stimuli presented *serially* should lead to similar deficits. This result has been reported following lesions in IT (Gaffan et al., 1986).

General Discussion

In this chapter I tested the hypothesis that lesion effects in PRh can be accounted for by assuming that PRh has visual information processing properties similar to those of other regions of inferotemporal cortex (IT). I constructed a neural network model of IT function that is based on two basic assumptions. First, neural circuitry intrinsic to IT is assumed to perform computations important for discrimination and perceptual learning of visual stimuli. Second, complex visual representations are assumed to be built up in a converging, hierarchical manner as information passes from primary visual areas, through IT to PRh. The Feature Conjunction Layer of the network, which corresponds to PRh, was lesioned and the lesioned and unlesioned network tested on discrimination paradigms that have

revealed PRh lesion-induced effects in monkeys. It was found that the foregoing simple assumptions appear to be all that is required to account for effects of lesions in PRh, including set size effects, impairments in "configural" learning, and greater effects on retention versus acquisition of visual discriminations. I thus present this model as a first step toward describing the possible neural mechanisms underlying "object identification" in PRh.

The Adaptive Value of Distributed Stimulus Representations

If objects are already fully specified by the representations of their features in early regions of IT, what is the advantage conferred by the complex representations in downstream regions such as PRh? The results of the present study suggest that complex representations disambiguate stimuli that have features in common, thus reducing interference. When attempting to solve complex visual discriminations, an animal with a dysfunctional PRh will rely more on the simple features to distinguish the stimuli. For such animals, discriminations between stimuli will be difficult in cases where there is much overlap between the features of the discriminanda (Experiments 1 and 2). As a result it will often be the case that the more difficult a discrimination problem is for an intact animal, the greater the impairment following a lesion in PRh. An exception to this rule would be expected, however, if the discrimination were made more difficult without manipulating stimulus features. Indeed, Hampton & Murray (1998) have shown that even though rotation of visual stimuli presented on a computer screen made discrimination more difficult for normal animals, monkeys with PRh lesions were not disproportionately impaired. (A similar result was found following lesions in IT; Gross, 1978 .) These results are consistent with the observation of rotational invariance in the response to specific stimuli of neurons in IT (e.g., Desimone et al., 1984; Perrett, Rolls, & Caan, 1982; Perrett et al., 1985). According to the present view, this is because object rotation—at least "isomorphic" rotation within the vertical plane—leaves features of the stimuli intact. Buckley & Gaffan (1998b), however, have reported that when three-dimensional objects are rotated *through* the vertical plane ("rotation-in-depth";

Gochin, 1996) PRh lesions impair discrimination. According to the present view this is because the stimulus will be *degraded* as certain features disappear behind the object, and other features are revealed as they move to the front of the object. This suggests a lower-level mechanism for the higher-level notion that PRh-lesioned animals appear to fail to recognize the rotated image *as the same object*. Importantly, such an account does not require separate representation of the 3-dimensionality of objects. Indeed, the 3-D structure of objects may be represented in the dorsal visual stream (Tanaka, 1996), suggesting a highly distributed representation of the complete set of attributes of an object (Tanaka, 1996).

Visual Information Processing versus Mnemonic Accounts of PRh Function

That lesion effects in PRh can be accurately simulated by assuming that PRh has visual information processing properties similar to those of IT leads to quite a different view of PRh function to that proposed by other authors. Specifically, Squire and Zola-Morgan (1991) suggest that PRh possesses mnemonic properties distinct from IT, and along with other structures such as the hippocampus, forms part of a unitary "medial temporal memory system", damage to any part of which can lead to impairments in "declarative", but not "procedural", memory. This view has, however, been challenged by experiments revealing functional double dissociations between lesions including PRh, and disconnections of the hippocampus via lesions of the fornix (Bussey et al., 1999; Ennaceur et al., 1996; Gaffan, 1994). Also consistent with the view that PRh has properties more in common with IT than with the hippocampus is evidence that lesions in PRh can lead to similar effects to lesions in other regions of IT on visual tasks including nonmatching-to-sample (Mishkin, 1982) and concurrent discrimination learning (Iwai & Mishkin, 1968; Malkova et al., 1995; Moss et al., 1981; Phillips et al., 1988). Moreover, neurons in PRh and in other areas of IT have similar properties, including firing patterns that may code for "recency" and "recognition" of visual stimuli (Fahy, Riches, & Brown, 1993). Finally, neuronal activity in the hippocampus has been shown to form a stable code for spatial position (O'Keefe & Nadel, 1978), whereas neuronal activity in PRh appears

to be a poor correlate of spatial position (Burwell, Shapiro, O'Malley, & Eichenbaum, 1998). Hippocampal neurons do not, however, appear to code the identity of objects. Thus I propose that in the absence of compelling evidence to the contrary, rather than possessing unique mnemonic functions distinct from those of IT, PRh may be best conceptualized as the final component of the ventral visual stream.

Stimulus information is encoded in the present model in PRh and endowed with meaning through associations with information in other brain regions. Thus this information is more akin to "semantic" or factual information than episodic memory which is associated with a specific context or learning episode. Indeed, there is evidence from human studies that impairments in semantic memory can result from lesions including damage to PRh (Hodges, Patterson, Oxbury, & Funnell, 1994; Kapur et al., 1994), and that temporal lobe lesions that spare PRh can lead to deficits in episodic, but not semantic memory (Kitchener, Hodges, & McCarthy, 1998; Vargha-Kadem et al., 1997). This is not to say that visual information in PRh could not be combined with spatial information from the hippocampus, for example, to form a complex episodic memory (Bussey et al., in press; Gaffan, 1994; Gaffan & Parker, 1996), merely that such interactions may not be obligatory (Bussey et al., in press).

The effects of lesions in PRh, according to the present model, are due not to the impairment of a particular type of learning or memory—for example, stimulus-reward or stimulus-response, declarative or procedural—but to compromising the representations of visual stimuli. Thus, *generally* stated, the present view is that the functions of PRh are more "perceptual" than "mnemonic". The perceptual learning achieved by the network does, however, result in the long-term memory of processed visual stimulus representations. In the model this corresponds to the adjusted weights of the connections to and from the Feature Layers and Feature Conjunction Layer, and in the association of this information with information in other brain regions. Because this information is not "consolidated" through, for example, its eventual storage elsewhere in the brain, it would be predicted that

lesions in PRh would not have a temporally-limited effect on retrograde memory—disruption of the storage of recently but not remotely learned object discrimination problems—like that observed following lesions of the hippocampus (e.g., Zola-Morgan & Squire, 1990). Indeed, Thornton, Rothblat and Murray (1997) found no evidence of a temporal gradient following lesions including PRh (although the use of more remote time points would be required to rule out the possibility of a more temporally extensive gradient). Furthermore, in humans damage extending beyond the hippocampus into temporal cortical regions including PRh can result in a temporally extensive, ungraded retrograde amnesia (Reed & Squire, 1998).

A Note on Delayed Matching and Nonmatching-To-Sample

It is now established that PRh is critically important for performance of the matching or nonmatching-to-sample tasks (dms; dnms), in which on a given trial the subject must report, during a choice phase, which of two objects had been seen during a previous sample phase (Meunier et al., 1993; Mumby & Pinel, 1994). Specifically, lesions including PRh lead to deficits on this task when large set-sizes are used (with very large stimulus sets the stimuli may be considered trial-unique), but have no effect when the task involves a small set of repeating stimuli (Eacott et al., 1994). A common interpretation of this pattern of results is that PRh is important for recognition memory (required when judging whether an object has been seen previously) as opposed to recency memory (required when judging which object has been seen most recently). Although the current model cannot operate in "real-time", and thus cannot simulate tasks such as dms and dnms, the principles of the model lead to a plausible, tentative reinterpretation of these results. First, the results of the simulations in the present study show that set-size effects in concurrent discrimination learning can be accounted for by assuming that PRh-lesioned animals rely on representations of simple features of stimuli to solve discriminations, and that lesion effects emerge when stimuli share a common set of features. Set size effects in dms and dnms might be explained in the same way, with inter- and intra-item interference in the large set-size version posing a particular problem for lesioned

animals. A related factor might be that the repeated presentations of the same stimuli in small set size dms and dnms may provide the opportunity to encode these stimuli in a number of different ways, and allow subjects to "sample" a number of different features on which the discrimination can be made. This would reduce the probability of feature overlap, thus aiding the lesioned animals. In contrast in the trial-unique version subjects are allowed to view the stimulus only once, reducing the opportunity for stimulus sampling and thus the number of features on which an animal can make the subsequent discrimination during the choice phase. This could lead to substantial difficulties for lesioned animals that must rely on an impoverished stimulus representation to make the discrimination. This "perceptual" versus "mnemonic" interpretation of PRh lesion-induced deficits on dms and dnms is supported by the observation that lesioned animals show deficits *even when no delay is interposed between sample and choice* (Eacott et al., 1994). To follow this line of inquiry still further, the delayed aspect of this task could provide additional difficulties for lesioned animals, as the decay of the stimulus trace during the delay could be thought of as reducing the feature representations on which the subsequent discrimination can be made, thus leading to impairments in lesioned animals (see discussion of stimulus degradation above). Indeed, the impairment in these tasks following PRh lesions is greatest when delays are longest. Furthermore, delays could lead to impairments in lesioned animals because in remembering a visual stimulus, it may be more efficient to keep one complex representation "on-line" in PRh than many representations of the component features. In this way, the combination of simpler representations into complex ones may be thought of as "visual chunking" similar to semantic chunking of verbal material (Miller, 1956). Indeed, Luck and Vogel (1997) have found evidence for such visual chunking in humans. This is of course highly speculative, and the issue is complicated as it has yet to be established by what mechanisms stimulus representations are held "on-line" during delays in dms and dnms, and whether such a putative working memory function is dependent on other brain regions such as the prefrontal cortex (Desimone, 1996). Nevertheless, these considerations suggest that the concept of object identification, perhaps via the mechanisms proposed in the

present model, may go some way toward accounting for PRh lesion effects on recognition memory tasks.

Conclusion

To summarize, I have constructed a model of IT function that is built on some very simple assumptions that are well-supported by the literature. By assuming that PRh is a downstream component of IT that contains representations of complex visual stimuli, this model can reproduce lesion effects in PRh that have been taken as evidence for a role for this region in "object identification". The extent to which a lesion in the model mimics effects of PRh lesions in animals suggests that there is no need to invoke special functions for PRh, distinct from those of other regions of IT, in order to account for the effects of lesions in PRh. Of course it would be premature to make the strong claim that special mnemonic or associative functions of PRh will *never* be revealed. Indeed, some evidence is now emerging that may point to such specialization of function. For example, it appears that PRh receives much greater brainstem dopaminergic input than other regions in IT (Lewis, Campbell, Foote, Goldstein, & Morrison, 1987), which could indicate a function for reward signals in this region, a notion at odds with the assumption that perceptual learning in PRh is unsupervised. This property of PRh may be related to the observation that some neurons in PRh appear to code for the temporal proximity of reward (Liu & Richmond, 1997), or that lesions which include PRh may disrupt object reversal learning (Bussey et al., in press; Murray, Baxter, & Gaffan, in press). As another example, PRh appears to receive many more polymodal inputs than other areas of IT (Suzuki & Amaral, 1994), and this could indicate a special function of PRh in, for example, cross-modal learning (Murray et al., 1998). Indeed, such polymodal information may allow the binding of visual and nonvisual features of objects (Murray et al., 1998). Another possibility is that the stimulus-stimulus associative functions of PRh could be important for associating different views of objects (Murray et al., 1998). It remains to be seen whether such putative associative functions are shared by other regions of IT. Finally, there is anatomical (Martin-

Elkins & Horel, 1992; Saleem & Tanaka, 1996), electrophysiological (Komatsu, Ideura, Kaji, & Yamane, 1992), and behavioral (Buckley et al., 1997; Horel, 1994) evidence that certain regions of IT (namely TEd) may be specialized for color processing and may comprise a color processing stream that does not rely critically on processing in PRh. Thus the present model may represent interactions between PRh and *TEv* only. The search for specialized functions of regions within IT therefore provides an important challenge for future research.

CHAPTER 7 CONCLUSIONS

The exploration of a problem in animal learning theory through a statistical model developed in the field of machine learning has contributed to, if not a completely new way of looking at the problem, at least a much more well-defined and testable theory. In addition, the use of the competitive learning technique to provide a concrete framework for the investigation of nonassociative differentiation has allowed me to show that certain experimental results that have until now been attributed to high-level attentional processes may be accounted for instead by simple, bottom-up differentiation. Third, the use of this connectionist framework has led to the beginning of work on defining the mechanisms and brain regions responsible for object identification. In the following section I list the specific contributions of this dissertation, and in the final section of this chapter I discuss some important issues and future research questions.

Contributions

The specific novel contributions of this dissertation research are enumerated below:

1. I have developed a novel connectionist model of perceptual learning that provides a mechanism for nonassociative differentiation (Gibson & Gibson, 1955). In contrast to other models (e.g.,McLaren, Kaye, & Mackintosh, 1989), in the current core model the mechanisms for these processes are compatible with a configural model of associative learning. The present model can account for critical perceptual learning phenomena such as exposure learning and effects of similarity on discrimination. It is also

shown that the model can explain the paradoxical result that preexposure to stimuli can either facilitate or impair subsequent discrimination learning.

2. I have provided an analysis of learning phenomena that are normally explained as being a result of "dimensional attention", and have shown that many of these phenomena may be more parsimoniously explained by the aforementioned core model of differentiation.

3. I have extended the core model to show how the integration of reinforcement and categorization information with the stimulus representation may contribute to the phenomena of acquired distinctiveness and equivalence.

4. I have also extended the core model to address issues regarding object representation in the brain; specifically, I have developed a model of the effects of lesions of monkey perirhinal cortex (PRh).

Further Questions

What is the relationship between top-down and bottom-up influences on stimulus representations?

In this dissertation it has been shown that a bottom-up, data-driven mechanism can account for many of the phenomena associated with perceptual learning. This is significant, as Occam's razor suggests that the best solution to a problem is the simplest, and the fewer mechanisms that are ascribed to an effect the better. However, there are effects that cannot be simulated by the current model, such as the ED/ID shift. Determining exactly what the difference between this effect and the effects that were accounted for by differentiation in Chapter 4 could yield insights into the attentional mechanisms that may be involved in perceptual learning. Furthermore, a future version of the current model in which a top-down attentional signal is incorporated could help to guide research on the regions of the brain responsible for this type of attention. Frontal cortex lesion studies of some of the

tasks simulated in Chapter 4 might also provide critical information regarding the relationship between top-down and bottom-up influences on stimulus representations and their loci in the brain.

Would a real-time model account for stimulus comparison effects?

The construction of a real-time version of the current model would enable the investigation of whether stimulus comparison effects can be accounted for within a differentiation framework, as discussed in Chapter 4. If so, this would be another subset of effects for which the current model could account, thereby providing more support for the idea that differentiation is the bottom-up mechanism of perceptual learning.

How plausible are multidimensional event representations?

Whereas the three-dimensional model presented in Chapter 5 may be psychologically reasonable, it is not particularly realistic physiologically. Moving on to a better representation for this idea would provide an additional constraint to the model and would enhance its plausibility

Is LVQ an alternate mechanism for acquired equivalence and distinctiveness effects?

The basic LVQ algorithm is the same as the standard Kohonen algorithm with two exceptions. First, each weight vector on the competitive layer has an associated class. Second, during training the weight update rule operates according to the class of the input vector. As in the standard algorithm, the winning competitive unit consists of the unit whose weights most closely match the input pattern. If the class of the competitive unit matches the class of the winning competitive unit, the competitive unit weights are updated as in the standard algorithm. If the classes of the units do not match, however, the competitive unit weights are updated to move them slightly away from the input pattern.. In order for LVQ to work, however, one would have to have spatially separated layers of similarly tuned units, one layer associated with class one and the other associated with class zero. If structured in

this manner, LVQ would be creating acquired equivalence and distinctiveness effects in essentially the same way as the extended model does, with stimuli associated with class one being represented relatively far away from stimuli associated with class zero and thus generalizing very little to each other.

Can memory effects such as spontaneous recovery and faster acquisition after extinction be explained by the extended model?

A completely different field of research involves the relationship between learning and memory in animals. As mentioned earlier, it is well known that reacquisition of a simple conditioning task after extinction occurs more quickly than did the initial acquisition (Konorski & Szwejkowska, 1950; Konorski & Szwejkowska, 1952). Interesting future work would involve the exploration of whether this group of memory-related effects fits into a multidimensional event representation framework.

How does the PRh model relate to configural learning theory?

The architecture of the PRh model in Chapter 6 is different from both configural and elemental learning models. It is essentially a hybrid of these two approaches in that both elements and configurations of elements are associated with the US. To this point, I have only looked at the model in relation to effects of PRh lesions. It would be interesting to apply the same model to experiments concerning learning in intact animals to see if this hybrid models makes predictions differing from those of straight configural and elemental theories.

CHAPTER 8 REFERENCES

Aggleton, J. P., Keen, S., Warburton, E. C., & Bussey, T. J. (1997). Extensive cytotoxic lesions involving both the rhinal cortices and area TE impair recognition but spare spatial alternation in the rat. Brain Res Bull, 43, 279-87.

Bennett, C. H., Wills, S. J., Wells, J. O., & Mackintosh, N. J. (1994). Reduced generalization following preexposure: Latent inhibition of common elements or a difference in familiarity? Journal of Experimental Psychology: Animal Behavior Processes, 20(3), 232-239.

Bennett, T. L., & Ellis, H. C. (1968). Tactual-kinesthetic feedback from manipulation of visual forms and nondifferential reinforcement in transfer of perceptual learning. Journal of Experimental Psychology, 77, 495-500.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. Psychological Review, 94, 115-147.

Booth, M. C. A., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cerebral Cortex, 8, 510-523.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. Psychological Bulletin, 114, 80-99.

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. Journal of Experimental Psychology: Animal Behavior Processes, 20, 219-231.

Brown, M. (1996). Neuronal responses and recognition memory. Seminars in the Neurosciences, 8, 23-32.

Buckley, M. J., & Gaffan, D. (1997). Impairment of visual object-discrimination learning after perirhinal cortex ablation. Behavioral Neuroscience, 111, 467-475.

Buckley, M. J., & Gaffan, D. (1998a). Learning and transfer of object-reward associations and the role of the perirhinal cortex. Behavioral Neuroscience, 112, 15-23.

Buckley, M. J., & Gaffan, D. (1998b). Perirhinal cortex ablation impairs visual object identification. Journal of Neuroscience, 18, 2268-2275.

Buckley, M. J., & Gaffan, D. (in press). Perirhinal cortex ablation impairs configural learning and paired-associate learning equally. Neuropsychologia.

Buckley, M. J., Gaffan, D., & Murray, E. A. (1997). Functional double dissociation between two inferior temporal cortical areas: Perirhinal cortex versus middle temporal gyrus. Journal of Neurophysiology, 77, 587-598.

Burwell, R., Shapiro, M., O'Malley, M., & Eichenbaum, H. (1998). Positional firing properties of perirhinal cortex neurons. NeuroReport, 9, 1013-1018.

Bussey, T. J. (1997). Rats with lesions of the fornix or rhinal cortex can solve the negative patterning task, Society for Neuroscience Abstracts (Vol. 23, pp. 1600).

Bussey, T. J., Duck, J., Muir, J. L., & Aggleton, J. P. (1999). Double dissociations resulting from two different hippocampal disconnections: A comparison of the behavioural effects of fornix transection with neurotoxic lesions of the perirhinal and postrhinal cortices in the rat. Manuscript submitted for publication .

Bussey, T. J., Muir, J., & Aggleton, J. (in press). Functionally dissociating aspects of event memory: The effects of combined perirhinal and postrhinal cortex lesions on object and place memory in the rat. Journal of Neuroscience.

Bussey, T. J., Warburton, E. C., Aggleton, J. P., & Muir, J. L. (1998). Fornix lesions can facilitate acquisition of the transverse patterning task: a challenge for "configural" theories of hippocampal function. J Neurosci, 18, 1622-31.

Carlson, J. G., & Wielkiewicz, R. M. (1972). Delay of reinforcement in instrumental discrimination learning in rats. Journal of Comparative and Physiological Psychology, 81, 365-370.

Carlson, J. G., & Wielkiewicz, R. M. (1976). Mediators of the effects of magnitude of reinforcement. Learning and Motivation, 7, 184-196.

Chamizo, V. D., & Mackintosh, N. J. (1989). Latent learning and latent inhibition in maze discriminations. Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 1, 21-31.

Channell, S., & Hall, G. (1981). Facilitation and retardation of discrimination learning after exposure to the stimuli. Journal of Experimental Psychology: Animal Behavior Processes, 7, 437-446.

Chantrey, D. F. (1972). Enhancement and retardation of discrimination learning in chicks after exposure to the discriminanda. Journal of Comparative and Physiological Psychology, 81, 256-261.

Couvillon, P. A., Tennant, W. A., & Bitterman, M. E. (1967). Intradimensional vs. Extradimensional transfer in the discriminative learning of goldfish and pigeons. Animal Learning and Behavior, 4, 197-203.

de Sa, V. R. (1994). Unsupervised classification learning from cross-modal environment structure. Unpublished Doctoral Dissertation, University of Rochester, Rochester, NY.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. Proceedings of the National Academy of Sciences of the USA, 93, 13494-13499.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci, 4, 2051-62.

Desimone, R., & Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. Brain Res, 178, 363-80.

Desimone, R., & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeys. In F. Boller & J. Grafman (Eds.), Handbook of Neuropsychology (Vol. 2, pp. 267-299): Elsevier Science.

Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. Nature, 380(6569), 69-72.

Dickinson, A. (1980). Contemporary animal learning theory. Cambridge: Cambridge University Press.

Dickinson, A. (1989). Expectancy theory in animal conditioning. In S. B. Klein & R. R. Mowrer (Eds.), Comtemporary learning theories: Pavlovian conditioning and the status of traditional learning theory (pp. 279-308). Hillsdale: Lawrence Erlbaum Associates.

Domjan, M., & Burkhard, B. (1986). The principles of learning and behavior. (2 ed.). Pacific Grove, CA: Brooks/Cole.

Eacott, M. J., Gaffan, D., & Murray, E. A. (1994). Preserved recognition memory for small sets and impaired stimulus identification for large sets following rhinal cortex ablations in monkeys. European Journal of Neuroscience, 6, 1466-1478.

Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. Behavioral and Brain Sciences, 17, 449-518.

Ennaceur, A., Neave, N., & Aggleton, J. P. (1996). Neurotoxic lesions of the perirhinal cortex do not mimic the behavioural effects of fornix transection in the rat. Behav Brain Res, 80, 9-25.

Espinet, A., Iraola, J. A., Bennett, C. H., & Mackintosh, N. J. (1995). Inhibitory association between neutral stimuli in flavor- aversion conditioning. Animal Learning and Behavior, 23(4), 361-368.

Estes, W. K. (1950). Toward a statistical theory of learning. Psychological Review, 57, 94-107.

Fahy, F., Riches, I., & Brown, M. (1993). Neuronal activity related to visual recognition memory: Long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. Experimental Brain Research, 96, 457-472.

Forgus, R. H. (1956). Advantage of early over late perceptual experience in improving discrimination. Canadian Journal of Psychology, 10, 147-155.

Frey, P. W., & Sears, R. J. (1978). Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. Psychological Review, 85(321-340).

Gaffan, D. (1994). Dissociated effects of perirhinal cortex ablation, fornix transection and amygdalectomy: Evidence for multiple memory systems in the primate temporal lobe. Experimental Brain Research, 99, 411-422.

Gaffan, D., Harrison, S., & Gaffan, E. (1986). Single and concurrent discrimination learning by monkeys after lesions of inferotemporal cortex. The Quarterly Journal of Experimental Psychology, 38B, 31-51.

Gaffan, D., & Murray, E. A. (1992). Monkeys (Macaca fascicularis) with rhinal cortex ablations succeed in object discrimination learning despite 24-hr intertrial intervals and fail at matching to sample despite double sample presentations. Behavioral Neuroscience, 106, 30-38.

Gaffan, D., & Parker, A. (1996). Interaction of perirhinal cortex with the fornix-fimbria: Memory for objects and "object-in-place" memory. Journal of Neuroscience, 16, 5864-5869.

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. Psychological Review, 47, 196-229.

Gibson, E. J. (1969). Principles of perceptual learning and development: New York: Appleton-Century Crofts.

Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. Journal of Comparative and Physiological Psychology, 49, 239-242.

Gibson, E. J., Walk, R. D., Pick, H. L., & Tighe, T. J. (1958). The effect of prolonged exposure to visual patterns on learning to discriminate similar and different patterns. Journal of Comparative and Physiological Psychology, 51, 584-587.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning -- differentiation or enrichment? Psychological Review, 62, 32-41.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. Hippocampus, 3, 491-516.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General, 123, 178-200.

Gonzalez, R. C., & Ross, S. (1958). The basis of solution by preverbal children of the intermediate-size problem. American Journal of Psychology, 71, 742-746.

Grice, G. R., & Davis, J. D. (1958). Mediated stimulus equivalence and distinctiveness in human conditioning. Journal of Experimental Psychology, 55, 565-571.

Grice, G. R., & Davis, J. D. (1960). Effect of concurrent responses on the evocation and generalization of the conditioned eyeblink. Journal of Experimental Psychology, 59, 391-395.

Grice, G. R., & Hunter, J. J. (1963). Response mediation of the conditioned eyeblink response. Journal of Experimental Psychology, 66, 338-346.

Gross, C. (1978). Inferior temporal lesions do not impair discrimination of rotated patterns in monkeys. Journal of Comparative and Physiological Psychology, 92, 1095-1109.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics, 23, 121-134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. Biological Cybernetics, 23, 187-202.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. Journal of Experimental Psychology, 51, 79-88.

Hall, G. (1991). Perceptual and associative learning. Oxford: Clarendon Press.

Hall, G., & Channell, S. (1985). A comparison of intradimensional and extradimensional shift learning in pigeons. Behavioural Processes, 10, 285-295.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. Journal of Experimental Psychology, 58, 321-333.

Hebb, D. O. (1949). The organization of behavior. New York: Wiley.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). Introduction to the theory of neural computation. (Vol. I). Redwood City, CA: Addison-Wesley.

Hinton, G. E., & Anderson, J. A. (Eds.). (1989). Parallel Models of Associative Memory (Second ed.): Lawrence Erlbaum Associates.

Hodges, J., Patterson, K., Oxbury, S., & Funnell, E. (1994). Semantic dementia: implications for the modularity of mind. Cognitive Neuropsychology, 11, 505-542.

Hogg, J., & Evans, P. L. (1975). Stimulus generalization following extra-dimensional training in educationally subnormal (severely) children. British Journal of Psychology, 66(2), 211-224.

Honey, R., & Hall, G. (1989). Enhanced discriminability and reduced associability following flavor preexposure. Learning and Motivation, 20, 262-277.

Honey, R. C., & Bateson, P. (1996). Stimulus comparison and perceptual learning: Further evidence and evaluation from an imprinting procedure. The Quarterly Journal of Experimental Psychology, 49B, 259-269.

Honey, R. C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. Quarterly Journal of Experimental Psychology, 47B, 83-103.

Horel, J. A. (1994). Retrieval of color and form during suppression of temporal cortex with cold. Behavioral Brain Research, 65, 165-172.

Horel, J. A., & Stegner, G. M. (1993). The effects of number of stimuli and prior exposure on performance of concurrent visual discriminations during suppression of inferotemporal cortex with cold. Behavioral Brain Research, 59, 161-168.

Hull, C. L. (1939). The problem of stimulus equivalence in behavior theory. Psychological Review, 46, 9-30.

Iwai, E., & Mishkin, M. (1968). Two visual foci in the temporal lobe of monkeys. In N. Yoshii & N. Buchwald (Eds.), Neurophysiological basis of learning and behavior (pp. 1-11). Japan: Osaka University Press.

Iwai, E., & Yukie, M. (1987). Amygdalofugal and amygdalopetal connections with modality-specific visual cortical areas in macaques (Macaca fuscata, M. mulatta, and M. fascicularis). Journal of Comparative Neurology, 261, 362-87.

Jagadeesh, B., & Desimone, R. (1997). Responses of inferior temporal (IT) neurons in the macaque monkey to morphed shapes, Society for Neuroscience Abstracts (Vol. 23, pp. 2229).

Jaynes, J. (1950). Learning a second response to a cue as a function of the magnitude of the first. Journal of Comparative and Physiological Psychology, 43, 398-408.

Jenkins, W. M., Merzenich, M. M., Ochs, M., Allard, T., & Guic-Robles, E. (1990). Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation. Journal of Neurophysiology, 63, 82-104.

Julesz, B. (1981). Textons, the elements of texture perception, and their interaction. nature, 290, 91-97.

Kamin, L. J. (1968). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), Miami symposium on the prediction of behavior: aversive stimulation (pp. 9-33). Miami: University of Miami Press.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (1991). Principles of Neural Science. (3 ed.). Norwalk, CT: Appleton and Lange.

Kapur, N., Ellison, D., Parkin, A., Hunkin, N., Burrows, E., Sampson, S., & Morrison, E. (1994). Bilateral temporal lobe pathology with sparing of medial temporal lobe structures: Lesion profile and pattern of memory disorder. Neuropsychologia, 32, 23-38.

Kasprow, W. J., Catterson, D., Schachtman, T. R., & Miller, R. R. (1982). Reminder-induced recovery of associations to an overshadowed stimulus. Learning and Motivation, 13, 308-318.

Kawachi, J. (1965). Effects of previous perceptual experience of specific three-dimensional objects on later visual discrimination behavior in rats. Japanese Journal of Psychological Research, 7, 20-27.

Kitchener, E., Hodges, J., & McCarthy, R. (1998). Acquisition of post-morbid vocabulary and semantic facts in the absence of episodic memory. Brain, 121, 1313-1327.

Klosterhalfen, S., Fischer, W., & Bitterman, M. E. (1978). Modification of attention in honey bees. Science, 201, 1241-1243.

Kobotake, E., Wang, G., & Tanaka, K. (in press). Effect of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. Journal of Neurophysiology.

Köhler, W. (1918). Nachweis einfarcher Strukturfunktionen beim Schimpansen und beim Haushuhn. Translated and condensed as "Simple structural functions in the chimpanzee and in the chicken". In W. D. Ellis (Ed.), A Source Book of Gestalt Psychology . London: Routledge and Kegan Paul.

Kohonen, T. (1982). Clustering taxonomy and topological maps of patterns. Paper presented at the Sixth International Conference on Pattern Recognition, Silver Springs, MD.

Kohonen, T. (1984). Self-organization and associative memory. Berlin: Springer-Verlag.

Komatsu, H., Ideura, Y., Kaji, S., & Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. Journal of Neuroscience, 12, 408-424.

Konorski, J. (1967). Integrative activity of the brain. Chicago: University of Chicago Press.

Konorski, J., & Szwejkowska, G. (1950). Chronic extinction and restoration of conditioned reflexes: I. Extinction against the excitatory background. Acta Biologiae Experimentalis, 15, 155-170.

Konorski, J., & Szwejkowska, G. (1952). Chronic extinction and restoration of conditioned reflexes: IV. The dependence of the course of extinction and restoration of conditioned reflexes on the "history" of the conditioned stimulus (The principle of the primacy of first training). Acta Biologiae Experimentalis, 16, 95-113.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.

Lawrence, D. H. (1949). Acquired distinctiveness of cues. I. Transfer between discriminations on the basis of familiarity with the stimulus. Journal of Experimental Psychology, 39, 770-784.

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. Journal of Comparative and Physiological Psychology, 45, 511-516.

Lawrence, D. H. (1963). The nature of a stimulus: some relationships between learning and perception. In S. Koch (Ed.), Psychology: a study of science (Vol. 5, pp. 179-212). New York: McGraw-Hill.

Lewis, D., Campbell, M., Foote, S., Goldstein, M., & Morrison, J. (1987). The distribution of tyrosine hydroxylase-immunoreactive fibers in primate neocortex is widespread but regionally specific. Journal of Neuroscience, 7, 279-290.

Liu, Z., & Richmond, B. (1997). Associative learning of task progress is coded by monkey perirhinal but not by TE neurons, Society for Neuroscience Abstracts (Vol. 23, pp. 1964).

Logan, F. A. (1966). Transfer of discrimination. Journal of Experimental Psychology, 71(4), 616-618.

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in primates. Cerebral Cortex, 3, 270-288.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: The effect of nonreinforced pre-exposure to the conditional stimulus. Journal of Comparative and Physiological Psychology, 52, 415-419.

Lubow, R. E., Schnur, P., & Rifkin, B. (1976). Latent inhibition and conditioned attention theory. Journal of Experimental Psychology: Animal Behavior Processes, 2, 163-174.

Luck, S., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.

Mackintosh, N. J. (1974). The psychology of animal learning.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review, 82(4), 276-298.

Mackintosh, N. J., Kaye, H., & Bennett, C. H. (1991). Perceptual learning in flavour aversion conditioning. Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 3, 297-322.

Mackintosh, N. J., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. Psychonomic Science, 14(1), 5-6.

Mackintosh, N. J., & Little, L. (1970). An analysis of transfer along a continuum. Canadian Journal of Psychology, 24(5), 362-369.

Malkova, L., Mishkin, M., & Bachevalier, J. (1995). Long-term effects of selective neonatal temporal lobe lesions on learning and memory in monkeys. Behavioral Neuroscience, 109, 212-226.

Martin-Elkins, C., & Horel, J. (1992). Cortical afferents to behaviorally defined regions of the inferior temporal and parahippocampal gyri as demonstrated by WGA-HRP. Journal of Comparative Neurology, 321, 177-192.

May, R. B., & MacPherson, D. F. (1971). Size discrimination in children facilitated by changes in task difficulty. Journal of Comparative and Physiological Psychology, 75(3), 453-458.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. (Vol. 2). Cambridge MA: MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5, 115-133.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), Parallel distributed processing: implications for psychology and neurobiology (pp. 102-130). Oxford: Clarendon Press.

Meador, K. J., Moore, E. E., Nichols, M. E., Abney, O. L., Taylor, H. S., Zamrini, E. Y., & Loring, D. W. (1993). The role of cholinergic systems in visuospatial processing and memory. Journal of Clinical and Experimental Neuropsychology, 15, 832-842.

Merzenich, M. M., Recanzone, G. H., Jenkins, W. M., & Grajski, K. A. (1990). Adaptive mechanisms in cortical networks underlying cortical contributions to learning and nondeclarative memory, Cold Spring Harbor Symposia on Quantitative Biology (Vol. LV,). Cold Spring Harbor: Cold Spring Harbor Press.

Meunier, M., Bachevalier, J., Mishkin, M., & Murray, E. A. (1993). Effects on visual recognition of combined and separate ablations of the entorhinal and perirhinal cortex in rhesus monkeys. Journal of Neuroscience, 13, 5418-5432.

Miller, E. K., Li, L., & Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. Science, 254, 1377-9.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Miller, R. R., & Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller & N. E. Spear (Eds.), Information processing in animals: Conditioned inhibition (pp. 51-88). Hillsdale, NJ: Erlbaum.

Mishkin, M. (1982). A memory system in the monkey. Philosophical Transactions of the Royal Society of London [Biology], 298, 83-95.

Moore, J. W., & Stickney, K. J. (1980). Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. Physiological Psychology, 8(207-217).

Moss, M., Mahut, H., & Zola-Morgan, S. (1981). Concurrent discrimination learning of monkeys after hippocampal, entorhinal, or fornix lesions. Journal of Neuroscience, 1, 227-240.

Mumby, D. G., & Pinel, J. P. J. (1994). Rhinal cortex lesions and object recognition in rats. Behavioral Neuroscience, 108, 11-18.

Murray, E., Baxter, M., & Gaffan, D. (in press). Monkeys with rhinal cortex damage or neurotoxic hippocampal lesions are impaired on spatial scene learning and object reversals. Behavioral Neuroscience.

Murray, E. A., & Bussey, T. J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. Trends in Cognitive Sciences, 3, 142-151.

Murray, E. A., Gaffan, D., & Mishkin, M. (1993). Neural substrates of visual stimulus-stimulus association in rhesus monkeys. Journal of Neuroscience, 13, 4549-4561.

Murray, E. A., Malkova, L., & Goulet, S. (1998). Crossmodal associations, intramodal associations, and object identification in macaque monkeys. In A. D. Milner (Ed.), Comparative Neuropsychology (pp. 51-67). New York: Oxford.

Neimark, E. D., & Estes, W. K. (1967). Stimulus sampling theory. San Francisco CA: Holden-Day.

Norcross, K. J. (1958). Effects of discrimination performance of similarity of previously acquired stimulus names. Journal of Experimental Psychology, 56, 305-309.

Norcross, K. J., & Spiker, C. C. (1957). The effects of type of stimulus pretraining on discrimination performance in preschool children. Child Development, 28, 79-84.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Oswalt, R. M. (1972). Relationship between level of visual pattern difficulty during rearing and subsequent discrimination in rats. Journal of Comparative and Physiological Psychology, 81, 122-125.

Pearce, J. M. (1987a). Introduction to animal cognition.

Pearce, J. M. (1987b). A model for stimulus generalization in Pavlovian conditioning. Psychological Review, 94, 61-73.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. Psychological Review, 101, 587-607.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87(6), 532-552.

Pearce, J. M., & Kaye, H. (1985). Strength of the orienting response during inhibitory conditioning. Journal of Experimental Psychology: Animal Behavior Processes, 11(3), 405-420.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stimulus on two discriminations. Journal of Experimental Psychology: Animal Behavior Processes, 19(2), 180-190.

Perrett, D., Rolls, E., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. Experimental Brain Research, 47, 329-342.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., & Jeeves, M. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. Proceedings of the Royal Society of London, 223, 293-317.

Peterson, G. B., & Trapold, M. A. (1980). Effects of altering outcome expectancies on pigeon's delayed conditional discrimination performance. Learning and Motivation, 11, 267-288.

Phillips, R. R., Malamut, B. L., Bachevalier, J., & Mishkin, M. (1988). Dissociation of the effects of inferior temporal and limbic lesions on object discrimination learning with 24-h intertrial intervals. Behavioral Brain Research, 27, 99-107.

Purtle, R. B. (1973). Peak Shift: A Review. Psychological Bulletin, 80, 408-421.

Redhead, E. S., & Pearce, J. M. (1995). Similarity and discrimination learning. Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 1, 46-66.

Reed, J., & Squire, L. (1998). Retrograde amnesia for facts and events: Findings from four new cases. Journal of Neuroscience, 18, 3943-3954.

Reese, H. W. (1972). Acquired distinctiveness and equivalence of cues in young children. Journal of Experimental Child Psychology, 13, 171-182.

Reid, L. S. (1953). The development of noncontinuity behavior through continuity learning. Journal of Experimetnal Psychology, 46, 107-12.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical Conditioning II: Current research and theory (pp. 64-99). New York: Appleton Century Crofts.

Roberts, A. C., Robbins, T. W., & Everitt, B. J. (1988). THe effects of intradimensional and extradimensional shifts on visual discrimination learning in humans and non-human primates. Quarterly Journal of Experimental Psychology, 40B, 321-341.

Roitblat, H. L. (1987). Introduction to comparative cognition. New York: Freeman.

Rothblat, L. A., Vnek, N., Gleason, T. C., & Kromer, L. F. (1993). Role of the parahippocampal region in spatial and non-spatial memory: Effects of parahippocampal lesions on rewarded alternation and concurrent object discrimination learning in the rat. Behav Brain Res, 55, 93-100.

Rumelhart, D. E., Hinton, G. E., & WIlliams, R. J. (1986). Learning internal representations by back-propagating errors. In D. E. Rumelhart & J. L. McClelland (Eds.), Parallel Distributed Processing (Vol. 1, pp. 318-362). Cambridge MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. (Vol. 1). Cambridge MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart & J. L. McClelland (Eds.), Parallel Distributed Processing (Vol. 1, pp. 151-193). Cambridge MA: MIT Press.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature, 354, 152-155.

Sakai, K., Naya, Y., & Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. Learning & Memory, 1, 83-105.

Saldanha, E., & Bitterman, M. (1951). Relational learning in the rat. American Journal of Psychology, 64, 37-53.

Saleem, K., & Tanaka, K. (1996). Divergent projections from the anterior inferotemporal area TE to the perirhinal and entorhinal cortices in the macaque monkey. Journal of Neuroscience, 16, 4757-4775.

Schmajuk, N. A., Lam, Y.-W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. Journal of Experimental Psychology: Animal Behavior Processes, 22, 321-349.

Schmajuk, N. A., & Moore, J. W. (1989). Effects of hippocampal manipulations on the classically conditioned nictitating membrane response: Simulations by an attentional-associative model. Behavioral Brain Research, 32(173-189).

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs, 75.

Spear, N. E. (1981). Extending the domain of memory retrieval. In R. R. Miller & N. E. Spear (Eds.), Information processing in animals: Memory mechanisms (pp. 341-378). Hillsdale, NJ: Erlbaum.

Spiker, C. C., & Norcross, K. J. (1962). Effects of previously acquired stimulus names on discrimination performance. Child Development, 33, 859-864.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. Science, 253, 1380-1386.

Sutherland, N. S., & Mackintosh, N. J. (1971). Mechanisms of animal discrimination learning. New York: Academic Press.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. Psychobiology, 17, 129-144.

Suzuki, W. A., & Amaral, D. G. (1994). Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. Journal of Comparative Neurology, 350, 497-533.

Symonds, M., & Hall, G. (1997). Stimulus preexposure, comparison, and changes in the associability of common stimulus features. The Quarterly Journal of Experimental Psychology, 50B, 317-331.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. Science, 262, 685-688.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annual Review of Neuroscience, 19, 109-139.

Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. Current Opinion in Neurobiology, 7, 523-529.

Tennant, W., & Bitterman, M. E. (1973). Some comparisons of intra- and extradimensional transfer in discriminative learning of goldfish. Journal of Comparative and Physiological Psychology, 83, 134-139.

Thornton, J. A., Malkova, L., & Murray, E. A. (1998). Rhinal cortex ablations fail to disrupt reinforcer devaluation effects in Rhesus monkeys (Macaca mulatta). Behavioral Neuroscience, 112, 1020-1025.

Thornton, J. A., Rothblat, L. A., & Murray, E. A. (1997). Rhinal cortex removal produces amnesia for preoperatively learned discrimination problems but fails to

disrupt postoperative acquisition and retention in rhesus monkeys. J Neurosci, 17, 8536-49.

Tovee, M. J., Rolls, E. T., & Ramachandran, V. S. (1996). Rapid visual learning in neurones of the primate temporal visual cortex. Neuroreport, 7, 2757-2760.

Trapold, M. A. (1970). Are expectancies based upon different positive reinforcing events discriminably different? Learning and Motivation, 1, 129-140.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.

Trobalon, J. B., Chamizo, V. D., & Mackintosh, N. J. (1992). Role of context in perceptual learning in maze discriminations. Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 1, 57-73.

Trobalon, J. B., Sansa, J., Chamizo, V. D., & Mackintosh, N. J. (1991). Perceptual learning in maze discriminations. Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 4, 389-402.

Ungerleider, L., Gaffan, D., & Pelak, V. (1989). Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. Experimental Brain Research, 76, 473-84.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), Analysis of visual behavior (pp. 549-586). Cambridge, MA: MIT Press.

Vargha-Kadem, F., Gadian, D., Watkins, K., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. Science, 277, 376-380.

Wagner, A. R. (1978). Expectancies and priming of STM. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), Cognitive processes in animal behavior (pp. 53-82). Hillsdale, NJ: Erlbaum.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In R. R. Miller & N. E. Spears (Eds.), Information processing in animals: Memory mechanisms (pp. 233-265). Hillsdale NJ: Erlbaum.

Walker, M. M., Lee, Y., & Bitterman, M. E. (1990). Transfer along a continuum in the discriminative learning of honeybees (Apis mellifera). Journal of Comparative Psychology, 104(1), 66-70.

Werbos, P. J. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. Unpublished Doctoral Dissertation, Harvard University, Boston, MA.

Zola-Morgan, S., & Squire, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. Science, 250, 288-290.

Zola-Morgan, S., Squire, L. R., & Ramus, S. J. (1994). Severity of memory impairment in monkeys as a function of locus and extent of damage within the medial temporal lobe memory system. Hippocampus, 4, 483-495.

APPENDIX A PSYCHOLOGICAL MODELS OF PERCEPTUAL LEARNING

Two different theories of perceptual learning dominate the literature. The first is fully situated within the associative framework, and suggests that perceptual learning is the result of associative processes operating on elemental stimulus representations (McLaren et al., 1989). The second suggests that perceptual learning is not related to associative learning, and that instead a nonassociative process called differentiation is responsible for the effects (Gibson & Gibson, 1955). The following sections outline the basic principles of each of these theories.

The associative view: McLaren, Kaye, and Mackintosh (1989)

A strength of the approach taken by McLaren et al. (1989) is its parsimony: they propose to account for preexposure effects by postulating only associative mechanisms and, furthermore, they account for one type of perceptual learning, exposure learning, with another type of perceptual learning, latent inhibition. Specifically, they suggest that a latent inhibition mechanism operating on stimulus elements might lead to the phenomena seen in exposure learning. While the authors' main assumption is that of stimulus sampling theory (Estes, 1950), they extend this by suggesting that in addition to associations between elements of different events, associations also form between elements comprising a single event, such as the CS itself. Variance in the stimuli is derived from the fact that only a subset of the elements is activated on each trial. Over time, associations will form between the most frequently sampled elements, thereby creating an elaborated stimulus representation that reflects the central tendency of the sampled elements. McLaren et

al. (1989) refer to this process as "unitization". The authors then argue that both latent inhibition and exposure learning stem from this associative elaboration of the stimulus representation over repeated presentations. They assume that latent inhibition is due to a decline in the associability of a repeatedly presented stimulus, which occurs as a consequence of the associations formed between stimulus elements and between stimulus and context elements during the presentations. Their postulated mechanism for this effect is a combination of Pearce and Hall's (1980) theory that associability decreases as the events following stimulus presentation become well predicted, and Wagner's (1981) theory which suggests that latent inhibition occurs because the representation of the stimulus becomes associated with the context in which it occurred, and is retrieved whenever the animal is placed in that context. An example of how their theory would operate is as follows. One case of exposure learning occurs when an animal is preexposed to stimuli *A* and *B*, then conditioned to *A* and later tested for generalization to *B* (as in the Hall and Honey experiments discussed in Chapter 3). According to McLaren and colleague's point of view, in this case conditioning occurs to the elements of *A* (*a*), some of which are common to *B* (*c*). Testing occurs with the elements of *B* (*b*) and the common elements *c*, therefore the extent of the generalization between *A* and *B* is due to the amount of conditioning to *c*. For non-exposed subjects, both *a* and *c* acquire associative strength according to their relative saliences. For preexposed subjects latent inhibition of the elements causes conditioning to *a* and *c* to proceed more slowly. Preexposure also facilitates discrimination of *A* and *B*, however, because *c* elements receive twice as much latent inhibition since they are present when either *A* or *B* is present. As the common elements are thus suppressed relative to the individual elements, later discrimination is facilitated. A second consequence of exposure to *A* and *B* will be the establishment of interconnections between their elements. This means that eventually the presence of a few *a* elements in a sample will retrieve others, thereby increasing the probability of correct choice on the trial. However, since associations will also form between *a* and *c* elements, this also suggests that exposure will increase generalization between the stimuli. McLaren et

al. (1989) avoid this problem by claiming that, since *a* and *b* elements will not be perceived exactly simultaneously, they will form *inhibitory* associations, which then cancel out the effects of the $a \rightarrow c$ associations and thus aid discrimination. In sum, according to McLaren and colleagues, exposure learning occurs as the result of three processes:

- The unitization process ensures that the presentation of a complex stimulus activates elements belonging to a central representation of the stimulus, rather than a small, variable subset of the elements.

- When two stimuli are preexposed, elements common to both, which are responsible for generalization between them, will be subject to a stronger latent inhibition effect than those unique to each.

- Associations between the common and unique elements of a stimulus will contribute to generalization, but this will be counteracted by the inhibitory associations that form between *a* and *b*.

The non-associative approach: Gibson and Gibson (1955)

Gibson (1940) originally argued that much of what needs to be learned in a paired associate task involves establishing a discrimination between the items. Performance improvement depends not just on the strengthening of the association between the stimulus and the response, but also on the extent to which the items tend to be confused. Gibson equated this process with a steepening of the gradient of generalization around each stimulus. Gibson called this phenomenon stimulus differentiation and assumed that it would transfer to and facilitate performance on a new task involving the same stimuli.

The crux of differentiation theory is that percepts change over time by an "elaboration of qualities, features, and dimensions of variation" (Gibson & Gibson, 1955 p. 34). With practice and experience a subject will become better able to detect distinctive features and distinguish one environmental event from another, and also become better at detecting invariant features that a given event displays from one occurrence to the next.

Gibson (1969) expands on these ideas by describing three processes that contribute to perceptual learning.

1. Abstraction is the process by which invariant relations are discovered over a number of objects or events. The relationship must be extracted through comparison of non-identical pairs, and does not necessarily reflect conscious search. The critical invariant may be independent of absolute stimulus values.

2. Filtering entails ignoring noncritical stimuli or idiosyncratic variations in stimuli.

3. Peripheral attentional mechanisms (e.g., fixation of the eyes) serve to expose sensory receptors to selected portions of the possible input.

The theory is intuitive and it captures many perceptual learning phenomena. Unfortunately, the mechanisms underlying the above processes were not outlined in detail beyond what is summarized above. As a result, the Gibsonian differentiation theory is little more than a restatement of the data. Furthermore, the specification is at such a high level that it is difficult to test the theory.

APPENDIX B PERIRHINAL CORTEX: A REGION AT THE INTERFACE BETWEEN PERCEPTION AND COGNITION

The perirhinal cortex (PRh) in the monkey is found at the ventromedial aspect of the temporal lobe, and is comprised of Brodmann areas 35 and 36 (see Figure C-1). It occupies the lateral bank of the rhinal sulcus plus a substantial portion of the inferior temporal gyrus. Recently, interest in PRh has increased dramatically, due in part to it being a region thought to be involved in (1) memory and (2) perception.

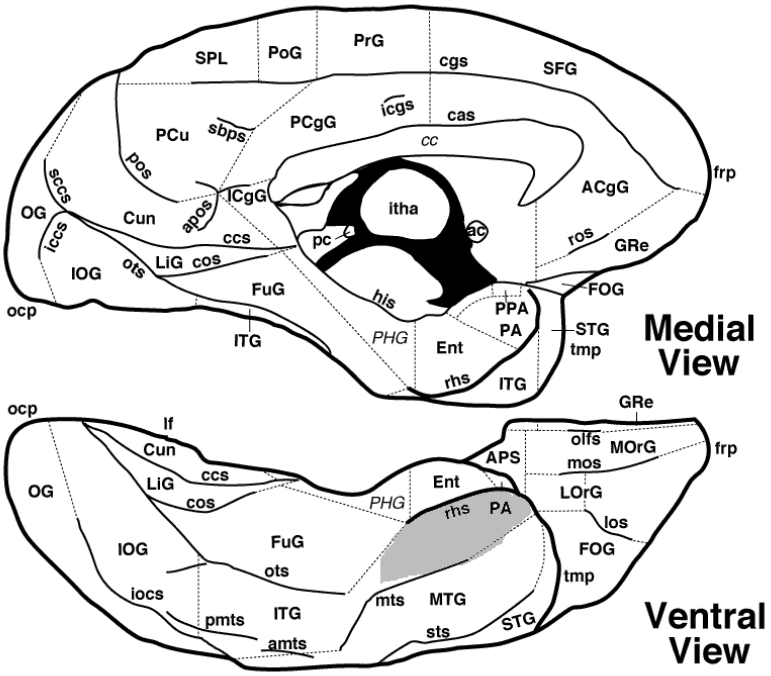


Figure C- 1: Medial and ventral views of the macaque brain. The approximate location of perirhinal cortex (visible only on the ventral view) is shaded gray (see Murray & Bussey, 1999).

(1) Memory. The current interest in the mnemonic functions of perirhinal cortex can be traced back to the case of the famous amnesic patient H.M.. H.M. became profoundly amnesic after receiving, as a treatment for intractable epilepsy, bilateral resection of medial temporal lobe structures including the hippocampus and amygdala. In order to understand better the cause of H.M.'s amnesia, as well as the neural basis of normal memory, researchers attempted to model amnesia by making combined lesions of the hippocampus and amygdala in monkeys, and testing these monkeys on a variety of memory tasks. One of the tasks that was consistently impaired by this lesion was the delayed nomatching-to-sample task, a measure of visual recognition memory. Later studies provided evidence, however, that the deficits observed for visual recognition memory were due not to damage to the hippocampus or the amygdala, but rather to inadvertent damage to a small strip of underlying cortex: the PRh. Thus it was established that PRh has an important role in recognition memory. Further investigations, however, revealed that lesions in this region impaired a wide variety of visual memory tasks, including paired associate learning (Murray, Gaffan, & Mishkin, 1993), concurrent discriminations (Buckley & Gaffan, 1998a), and configural learning (Buckley & Gaffan, in press). Thus it seems that the role of PRh cortex is clearly not limited to recognition, but has a more general role in what some workers have called "object identification" (Buckley & Gaffan, 1998b; Murray et al., 1993).

(2) Perception. PRh has also been of great interest to researchers interested not in the neural basis of memory *per se*, but in what might be called "perception": the basic aspects of visual information processing. Most notably, Ungerleider (personal communication) now considers PRh to be a part of the well-known "ventral visual stream" or "what" visual pathway (Ungerleider & Mishkin, 1982).

These two views do not appear to be mutually exclusive. Indeed, much of the evidence arguing for inclusion of PRh in the ventral visual stream comes from the aforementioned lesion studies of object recognition. Prominent and influential views of the neural substrates of memory and perception, however, continue to segregate

these two functions in the brain. Specifically, it has been proposed that whereas PRh has a role exclusively in memory, regions of IT upstream from PRh (e.g., area TE) have a role exclusively in perception. The model presented in Chapter 6 may go some way toward reconciling these views, showing how PRh could be thought of as important for *both* memory and perception.