

## **Chapter 2: Qualitative Model Comparisons**

Under ideal circumstances, researchers would like to base their comparison of models upon empirical tests of their basic principles, free from dependence on any specific ad hoc assumptions. Furthermore, it is best if the models being compared make different predictions that do not depend on any specific parameter values, but rather hold true for any set of parameter values. This chapter examines mathematical and computational tools that can be used to perform these types of qualitative tests of cognitive models (see Roberts & Pashler, 2000, for more discussion of this issue).

Several highly sophisticated methods have been developed for qualitatively testing models. Fundamental measurement theorists (cf., Krantz, Luce, Suppes, Tversky, 1972; Wallsten, 1978) have derived basic axioms for testing judgment and decision theories. Information processing theorists (Colonious & Marley, 19xx; Townsend and Schwieckert, 19xx) have derived critical properties for distinguishing serial and parallel information processing systems. Stochastic choice theorists have derived general conditions for comparing model probabilistic models of preferential choice (Luce & Suppes, 1965; Tversky, 1972). Recently, Dunn (in press) has proposed a general method, called the signed difference test, for qualitatively testing a wide range of cognitive models. This chapter will provide an example using a method from fundamental measurement theory.

The chapter is organized as follows. First, we describe a fictitious experiment on category learning that examines performance of amnesic and normal participants on

categorization and recognition tasks, and we summarize the results of this fictitious experiment. Second, two popular cognitive models of category learning are presented – one is a connectionist version of a prototype model, and the other is a connectionist version of an exemplar model. Third, a qualitative comparison of the two models is presented that employs a test based on fundamental measurement theory. In the final section, a qualitative analysis is used to determine whether it is possible to explain the results using a single memory system, or whether it is necessary to posit multiple memory systems.<sup>1</sup>

### **1. Category Learning Experiment.**

*Stimuli.* Category learning is a major topic in Cognitive Science, and human category learning has been studied in the experimental laboratory for almost a century. A typical experiment consists of creating different sets of artificial stimuli. One reason for using artificial rather than natural stimuli is that the participants of the experiment have no prior knowledge or experience of the categories. Another advantage of using artificial stimuli is that the experimenter gains a great deal of experimental control over the stimuli. As we shall see later, this is crucial for the qualitative test of the models that we wish to perform.

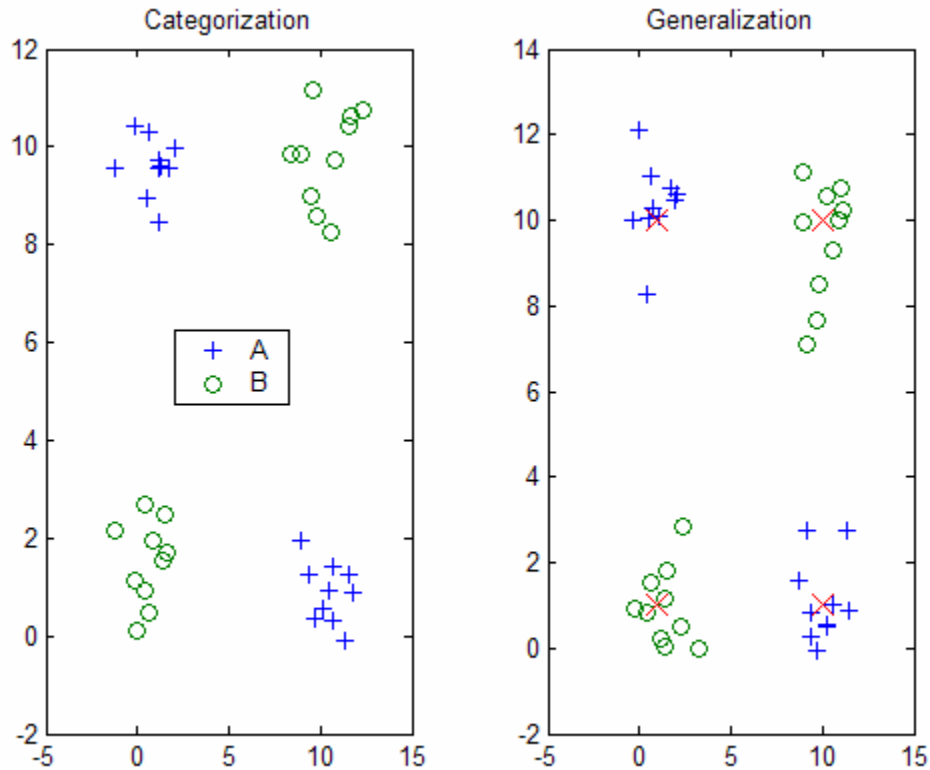
For our purposes, imagine that the stimuli are artificially generated X – ray images showing a dangerous looking node that may be cancerous or benign. In many category experiments, the stimuli (e.g. X – ray images) are designed to vary primarily as a function of two variables – for example, the diameter and brightness of a node in an X-

---

<sup>1</sup> The design of the fictitious experiment described here was inspired by Knowlton and Squires (1993) research, and the analysis of single versus dual memory systems was inspired by the work of Nosofsky and Zaka (1998). However, the design and results presented here were modified to meet the pedagogical needs of this chapter.

ray image. In this case, the stimuli can be represented as points in a two dimensional space. Figure 1, shown below, illustrates the stimuli used in the fictitious experiment presented in this chapter. There are two panels shown in Figure 1, the left represents stimuli presented during training, and the right represents stimuli presented later during a transfer test. The horizontal axis of each panel represents variation along the first stimulus dimension (e.g., diameter of a node), and the vertical axis represents the variation along the second dimension (e.g., brightness of a node). Each point represents a stimulus, with points marked by a plus sign representing stimuli belonging to category A, and points marked by a circle representing stimuli belonging to category B. Category A stimuli tend to be either low on the first dimension and high on the second dimension, or they tend to be high on the first dimension and low on the second dimension. Category B stimuli tend to be either low on both dimensions, or high on both dimensions. This kind of arrangement between stimuli and categories is called an exclusive-or (XOR) problem (see Nosofsky, 19xx, for a real study using these types of stimuli).

Figure 1: Stimuli used in the Category Learning Experiment



*Procedure.* Usually there are two phases to a category learning experiment, a training phase followed by a transfer test phase. During the training phase of the experiment, each participant is presented a stimulus (e.g., X-ray image) and asked to categorize the stimulus into one of two categories: category A (e.g., a cancerous node) versus category B (e.g., a benign node). After making each category decision, the participant is given feedback indicating the correct category assignment. In a typical experiment, each participant is trained on the ensemble of stimuli for several replications. In this case, we simulated 10 repetitions, producing a total of  $20 \text{ (stimuli)} \times 2 \text{ (categories)} \times 10 \text{ (repetitions)} = 400$  training trials per person.

Following this training, each participant is given one of two types of transfer tests – either a generalization test or a recognition test. No feedback is presented during the

transfer phase, and this phase is designed to assess what has been learned after training is completed.

For the generalization test, the participant is asked to categorize all the stimuli shown in both panels. The stimuli in the right panel, which were not presented during training, are called generalization test stimuli. There are 20 such stimuli generated from each of the two categories. In addition, there are four special transfer stimuli marked by the red crosses shown in the right panel. Note that these special transfer stimuli lie at the centroid of each stimulus cluster, and they are located at dimension values [1,1], [1,10], [10,1] and [10,10]. These special transfer stimuli are critical for the qualitative model comparison presented later.

The second type of transfer test is a recognition test. In this case, the participant is asked to decide whether the presented stimulus is a new transfer stimulus (never seen before) or an old stimulus experienced during training. In this condition, the participant makes an old versus new decision to all of the stimuli shown in both panels.

*Participants.* Suppose that two different populations of participants are recruited for the experiment: One is a group of 50 amnesic individuals that suffer severe episodic memory deficits, and the second is a group of 50 individuals with normal memory. The amnesic group is expected to perform very poorly on the recognition test. The critical question is – how well does the amnesic group perform on the categorization task? Assume that the groups are approximately matched with respect to age. (See Knowlton and Squire, 1993, for a real study of these two populations).

*Transfer Test Results.* Table 1 contains the (fictitious) probabilities of choosing category A for the four special transfer stimuli (the red crosses at the centroid of each

stimulus cluster in the right panel of Figure 1), averaged across all participants. Note that this table reveals a crossover interaction effect: when the first dimension was fixed at a high value (first row), the probability of choosing category A decreased as the value of the second dimension increased; however, when the first dimension was fixed at a low value (second row), the probability of choosing category A increased as the value of the second dimension increased. Thus the value of the second dimension had opposite effects on response probability, depending on the value of the first dimension. These results are exactly what one would expect if the participants accurately learned the category assignments for each cluster of stimuli. These results are critical for the model comparisons discussed later.

	$s_2 = 1$	$s_2 = 10$
$s_1 = 10$	.94	.03
$s_1 = 1$	.05	.98

Table 2 presents the results for the transfer test stimuli broken down by type of type of transfer test and type of group. The column labeled ‘Proportion Correct Categorization’ shows the proportion of the correctly categorized stimuli, pooled across the 40 generalization test stimuli (not including the four special transfer stimuli), and averaged across the participants within each group. The column labeled ‘ $d'$  Recognition’ is a commonly used measure of recognition performance: if we define  $h = \text{Pr}[\text{respond old} \mid \text{training stimulus}]$  and  $f = \text{Pr}[\text{respond old} \mid \text{transfer stimulus}]$  then

$$d' = \ln[h/(1-h)] - \ln[f/(1-f)].$$

(The rationale behind this measure is explained in the next section after we describe the model for recognition responses). The results shown in the table were obtained by first computing  $d'$  separately for each participant, and then averaging across participants within each group.

Group	Proportion Correct Categorization	$d'$ Recognition
Normal Group	.95	1.26
Amnesic Group	.98	.15

Table 2 shows a puzzling interaction effect – as expected, the normal group performed much better than the amnesic group on the recognition test, and in fact the recognition performance of the amnesic group is near zero. Surprisingly, the amnesic group performed about equal or slightly better than the normal group on the generalization test. The amnesic group seemed capable of performing the categorization task at transfer without being able to recognize the old training stimuli.

These results lead one to think that perhaps categorization and recognition are based on separate memory systems (see Knowlton and Squire, 1993). One is an implicit system (e.g., prototype learning) used for categorization, and another is an explicit system (e.g., exemplar learning) used for recognition. Both systems are effective for the individuals with normal memory. The implicit system remains intact but the explicit system is damaged for the amnesic individuals.

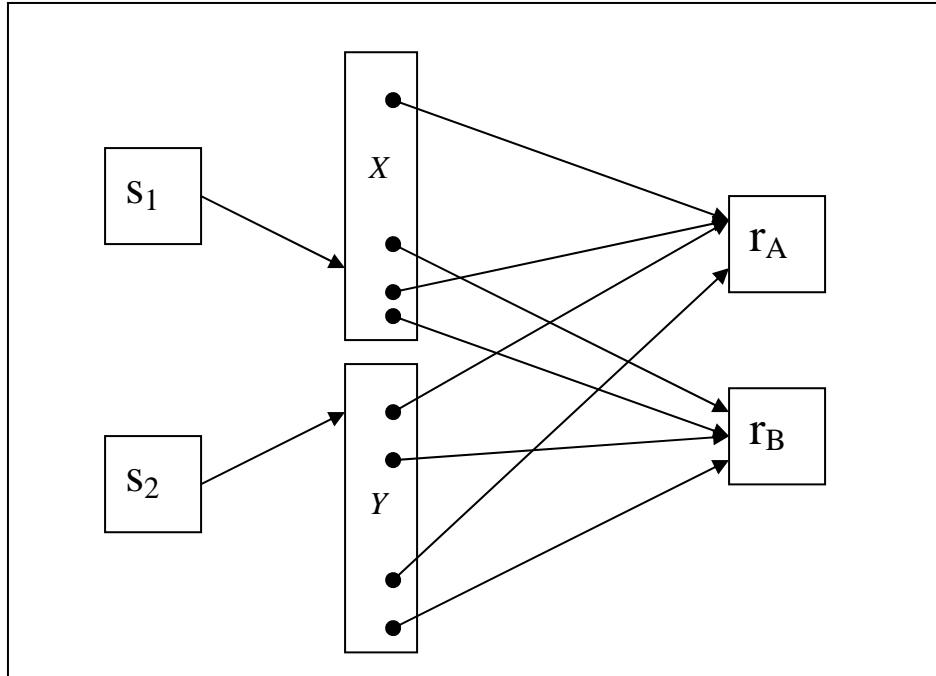
## 2. Two Models of Category Learning.

There a large number of sophisticated models for category learning in the literature (for some recent versions see Nosofsky & Palmeri, 1997; Ashby & Maddox, 19xx; Kruschke, 1992; Gluck & Bower, 1992). We will focus on two reasonably simple connectionist types of models to illustrate methods for qualitative comparisons of models. One model is called a connectionist version of a prototype model, and another model is called a connectionist version of an exemplar model.

To present these models, we need to introduce some notation. The stimulus presented on training trial  $t$  is denoted  $S(t) = [s_1(t), s_2(t)]$  where  $s_1(t)$  represents the value of the stimulus on the first dimension and  $s_2(t)$  represents the value of the stimulus on the second dimension. For example, if the special transfer indicated by the red cross in the lower right hand corner of Figure 1b is presented, then  $S(t) = [10, 1]$ . When a stimulus from category A is presented on trial  $t$ , then the target feedback is represented numerically by  $F(t) = [f_1(t), f_2(t)]$ , where  $F(t) = [1, 0]$  if a stimulus from category A is presented on trial  $t$ , and  $F(t) = [0, 1]$  if a stimulus from category B is presented on trial  $t$ .

*Prototype Model.* The prototype model is illustrated in Figure 2 shown below. Briefly, this model is based on three sets of assumptions. First, there are two sets of inputs nodes: One set (labeled  $X$ ) is activated by the value of the stimulus on the first dimension ( $s_1$ ), and the other set (labeled  $Y$ ) is activated by the value of the stimulus on the second dimension ( $s_2$ ). Second, these inputs are passed through a set of weighted connections to the output nodes corresponding to each category, which then generate the response. Finally, the connection weights are updated on the basis of feedback following each response during training. The details about the assumptions are presented next.

Figure 2: A Connectionist Version of a Prototype Model



The prototype model assumes that two sets of input nodes are used to represent the stimulus,  $S(t)$ . The number of nodes in each set is denoted by  $m$ , and so there are a total of  $2 \cdot m$  nodes. The first set of nodes are designed to detect values of the stimulus on the first dimension,  $s_1(t)$ ; and the second set of nodes are designed to detect values of the stimulus on the second dimension,  $s_2(t)$ . Each node within a set is designed to detect a particular stimulus value, which is called the ideal point of the node. The  $i$ -th node in the first set is designed to detect a stimulus value denoted  $X_i$ , and the activation of this node, denoted  $x_i(t)$ , is determined by the similarity of  $s_1(t)$  to  $X_i$ . The  $i$ -th node in the second set is designed to detect a stimulus value denoted  $Y_i$ , and the activation this node, denoted  $y_i(t)$ , is determined by the similarity of  $s_2(t)$  to  $Y_i$ . The similarity between the current stimulus value and the ideal point for each node is determined by a Gaussian type of generalization gradient

$$\text{sim}(X_i, s_1) = e^{-\left(\frac{|X_i - s_1|}{\sigma}\right)^2} \quad (1a)$$

$$sim(Y_i, s_2) = e^{-\left(\frac{Y_i - s_2}{\sigma}\right)^2} \quad (1b)$$

The parameter,  $\sigma$ , in the above equations, is called the discriminability parameter, and it determines the width or spread of the activation around the ideal point. A low discriminability parameter (large  $\sigma$ ) makes it hard to discriminate differences between stimulus values, and a high discriminability parameter (small  $\sigma$ ) makes easy to discriminate differences between stimulus values.

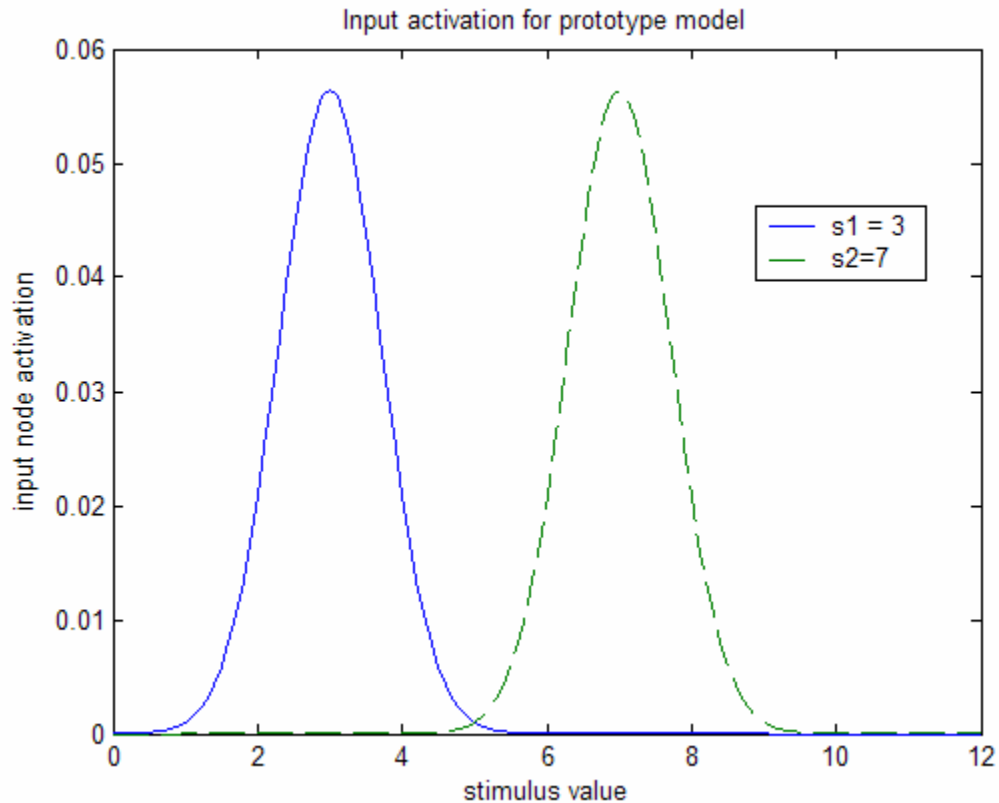
The input activation generated at the  $i$ -th node is determined by the similarity of that node relative to the similarity of all the nodes:

$$x_i(t) = \frac{sim(X_i, s_1)}{\sum sim(X_i, s_1)},$$

$$y_i(t) = \frac{sim(Y_i, s_2)}{\sum sim(Y_i, s_2)}.$$

A stimulus produces a distribution of activation across the input nodes for each set, resulting in two separate distributions. The distribution for each set is centered on the stimulus value for the corresponding dimension. Figure 3, shown below, shows the two distributions produced by setting  $S(t) = [3,7]$ . To generate this figure, we used  $m = 121$  equally spaced nodes with ideal points covering the stimulus range from 0 to 12, and we set the discriminability parameter equal to  $\sigma = 1$ . Note that the distribution for the first set is centered on  $s_1 = 3$  and the distribution for the second set is centered on  $s_2 = 7$ .

Figure 3: Input activation for the prototype model.



To see more concretely how this input activation function works, let us take a very simple example. Suppose that we use only  $m = 3$  equally spaced input nodes to cover the range of our stimuli. One node is designed to detect low stimuli, a second is designed to detect intermediate stimuli, and a third is designed to detect large stimuli. The stimuli range in value from 0 to 12 so we set ideal points for the three detectors equal to  $X_1 = 0$ ,  $X_2 = 6$ , and  $X_3 = 12$  for the first set; and  $Y_1 = 0$ ,  $Y_2 = 6$ , and  $Y_3 = 12$  for the second set. Suppose the discriminability parameter is equal to  $\sigma = 3$ . Also suppose that the stimulus,  $S(t) = [3, 7]$  is presented. First we compute the similarity between the stimulus value  $s_1(t) = 3$  and the three nodes for the first set of input nodes:

$$e^{-\left(\frac{[0-3]}{3}\right)^2} = .368, \quad e^{-\left(\frac{[6-3]}{3}\right)^2} = .368, \quad e^{-\left(\frac{[12-3]}{3}\right)^2} = .000,$$

These similarities sum up to  $.368 + .368 + .000 = .736$ , and the resulting activations are

$$x_1(t) = .368/.736 = .50, x_2(t) = .368/.736 = .50, x_3(t) = .000/.736 = .00.$$

We do this again for the second set using the stimulus value  $s_2(t) = 7$ :

$$e^{-\left(\frac{10-7}{3}\right)^2} = .004, e^{-\left(\frac{16-7}{3}\right)^2} = .895, e^{-\left(\frac{112-7}{3}\right)^2} = .062.$$

The sum of similarities are .961 for the second set, and so the resulting activations are

$$y_1(t) = .004/.961 = .004, y_2(t) = .895/.961 = .931, \text{ and } y_3(t) = .062/.961 = .065.$$

This same procedure was used to generate Figure 2, except that we used 121 nodes rather than just 3 nodes.

The input nodes are connected to two category nodes, one for each category. The activation of the two category nodes are denoted  $r_A(t)$  and  $r_B(t)$  for category A and B, respectively. The connection weight,  $v_{ik}(t)$ , connects the input activation  $x_i(t)$  to the  $k$ -th category output; the connection weight,  $w_{ik}(t)$ , connects the input activation  $y_i(t)$  to the  $k$ -th category output. The activation of the category nodes is based on the following linear input to output mapping<sup>2</sup>:

$$r_k(t) = \sum_i v_{ik}(t) \cdot x_i(t) + \sum_i w_{ik}(t) \cdot y_i(t) . \quad (2)$$

This model is called a prototype model because the set of weights,  $[v_{1A}, \dots, v_{mA}; w_{1A}, \dots, w_{mA}]$  connecting the inputs to the output for category A forms a representation of the prototype pattern for category A. The more similar the input activation pattern is to these weights, the more likely the stimulus matches the prototype for category A. Likewise, the set of weights,  $[v_{1B}, \dots, v_{mB}; w_{1B}, \dots, w_{mB}]$  connecting the inputs to the output for category B forms a representation of the prototype input pattern for category B.

---

<sup>2</sup> Some connectionist models (see Rumelhart & McClelland, 1986) include an extra layer of nodes that perform a nonlinear mapping of activation. However, many applications to category learning data (Gluck & Bower, 1992; Kruschke, 1992) use linear mappings between inputs and outputs.

To be more concrete, reconsider the earlier example with only three input nodes for each stimulus dimension. In this case, there are three connection weights connecting the first set of nodes to category A, and another three connection weights connecting the second set of nodes to category A. The same holds for category B, producing a total of 12 connection weights. Suppose after some amount of training, that the connection weights from the first set of nodes to category A are [ $v_{1A} = 1, v_{2A} = .02, v_{3A} = .01$ ], and the connection weights from the second set of nodes to category A are [ $w_{1A} = .02, w_{2A} = .01, w_{3A} = 1$ ]. In this case, the prototype for category A is a low value on the first dimension and a high value on the second dimension. Suppose the connection weights from the first set to category B are [ $v_{1B} = .02, v_{2B} = 1, v_{3B} = .01$ ], and the connection weights from the second set of nodes to category B are [ $w_{1B} = .01, w_{2B} = 1, w_{3B} = .02$ ]. Then the category B prototype has intermediate values on both dimensions. If the stimulus  $S(t) = [3,7]$  is presented, then using the input activation values computed earlier, the output activation to category A is

$$r_A = [ (1)(.500) + (.02)(.500) + (.01)(.000) ] \\ + [ (.02)(.004) + (.01)(.931) + (1)(.065) ] = .5844$$

The output activation to category B is

$$r_B = [ (.02)(.500) + (1)(.500) + (.01)(.000) ] \\ + [ (.01)(.004) + (1)(.931) + (.02)(.065) ] = 1.4373$$

Therefore, the input activation pattern produced by  $S(t) = [3,7]$  matches the category B prototype better than it matches the category A prototype.

The connection weights are updated according to an error reduction or delta learning rule (see Stone, 1986, for further discussion):

$$v_{ik}(t+1) = v_{ik}(t) + \alpha \cdot [f_k(t) - r_k(t)] \cdot x_i(t) \quad (3a)$$

$$w_{ik}(t+1) = w_{ik}(t) + \alpha \cdot [f_k(t) - r_k(t)] \cdot y_i(t) \quad (3b)$$

The delta rule is based on the following simple idea. The difference  $[f_k(t) - r_k(t)]$  is called the error signal because it equals the difference between the observed feedback and the prediction of that feedback. For example, if category A is the correct stimulus so that  $f_A = 1$ , and if  $r_A(t) = .5$ , which is too low so that the error is positive, then the weight is increased making the activation stronger upon the next appearance of this stimulus.<sup>3</sup>

To be more concrete, reconsider the earlier example where we computed the category outputs to the stimulus  $S(t) = [3,7]$ . In that case, the category A output was  $r_A = .584$ , and the category B output was  $r_B = 1.4373$ . Suppose the correct category for this stimulus is B, in which case  $f_2(t) = 1$  for this stimulus. In this case, the output response exceeds the feedback value, and so the error is negative, indicating that the weight needs to be decreased. Consider the change in the weight,  $w_{2B}(t)$ , connecting the activation,  $x_2(t)$ , to the category B response. Recall that the connection weight used to compute the output in this example was  $w_{2B}(t) = 1.00$ . Suppose the learning rate is  $\alpha = .25$  and recall that  $x_2(t) = .50$  in this example. Then this weight is updated for the next trial as follows

$$w_{2B}(t+1) = 1.00 + (.25)(1 - 1.4373) \cdot (.50) = .9453,$$

which is a change in the appropriate direction. This same procedure is applied to all 12 connection weights after each feedback trial.

For a categorization task, the probability of choosing category A is based on a ratio of strength of the output activations. After  $t$  trials of training

---

<sup>3</sup> The delta learning rule can be derived from the gradient of the sum of squared prediction errors with respect to the connection weights (see Stone, 1986).

$$\Pr[A | S(t)] = \frac{e^{b \cdot r_A(t)}}{e^{b \cdot r_A(t)} + e^{b \cdot r_B(t)}} \quad (4)$$

The coefficient,  $b$ , is called a sensitivity parameter, which determines the sensitivity of choice probability to the activation of each category. Increasing the sensitivity parameter increases the slope of the function relating choice probability to the activation of a category. The ratio of strength choice rule is commonly used in connectionist models of choice. This choice rule originated from a theory of choice proposed by Luce (1959).

Reconsidering the previous example once again, recall that when the stimulus  $S(t) = [3,7]$  was presented, then the outputs to categories A and B were  $r_A = .584$  and  $r_B = 1.4373$ . In this case, category B is the favored response. If we set  $b = 1$  then the probability of incorrectly choosing category A equals

$$\frac{e^{(1) \cdot (.584)}}{e^{(1) \cdot (.584)} + e^{(1) \cdot (1.4373)}} = .2987,$$

which is low but there is still a fair chance for making an error. However, if we increase this to  $b = 2$ , then the probability of error drops to .1536. Thus as the parameter,  $b$ , increases, the probability of error decreases.

For a recognition task, the probability of making an old recognition response is assumed to be increasing function of the total amount of activation produced by the stimulus to both output nodes (cf. Nosofsky & Zaka, 1998). More specifically, a logistic function is used to related total activation to old-new recognition response probability:

$$\Pr[old | S(t)] = \frac{e^{c \cdot [r_A(t) + r_B(t)]}}{\beta + e^{c \cdot [r_A(t) + r_B(t)]}} \quad (5a)$$

$$\Pr[new | S(t)] = 1 - \frac{e^{c \cdot [r_A(t) + r_B(t)]}}{\beta + e^{c \cdot [r_A(t) + r_B(t)]}} = \frac{\beta}{\beta + e^{c \cdot [r_A(t) + r_B(t)]}} \quad (5b)$$

The parameter  $c$  determines the sensitivity of the recognition probability to the category activations. Increasing the sensitivity parameter,  $c$ , causes the recognition probability to be more strongly influenced by the category activations.

The parameter  $\beta$  is used to represent the activation favoring a new response produced by the associations of the test stimulus with background context features. The background context is assumed to be constant across trials, and therefore unrelated to new or old stimuli, and so this parameter does not depend on whether the test stimulus is new or old. Thus it is response bias parameter representing the tendency to say new to any stimulus, and increasing  $\beta$  increases the tendency to respond new.

The  $d'$  index, commonly used to measure recognition performance, is designed to be insensitive to this response bias parameter. To see this, define  $h = \Pr[\text{respond old} \mid \text{training stimulus}]$ ; and define  $r_{\text{old}}$  as the summed output,  $r_A(t) + r_B(t)$ , that is generated by an old training stimulus. Then from Equation 5 we find that

$$\frac{h}{1-h} = \frac{e^{c \cdot r_{\text{old}}}}{\beta} \quad \text{and} \quad \ln\left(\frac{h}{1-h}\right) = c \cdot r_{\text{old}} - \beta.$$

In a similar manner, define  $f = \Pr[\text{respond old} \mid \text{transfer stimulus}]$ ; and define  $r_{\text{new}}$  as the summed output,  $r_A(t) + r_B(t)$ , that is generated by a new transfer stimulus. Then from Equation 5 we find that

$$\frac{f}{1-f} = \frac{e^{c \cdot r_{\text{new}}}}{\beta} \quad \text{and} \quad \ln\left(\frac{f}{1-f}\right) = c \cdot r_{\text{new}} - \beta.$$

The difference yields

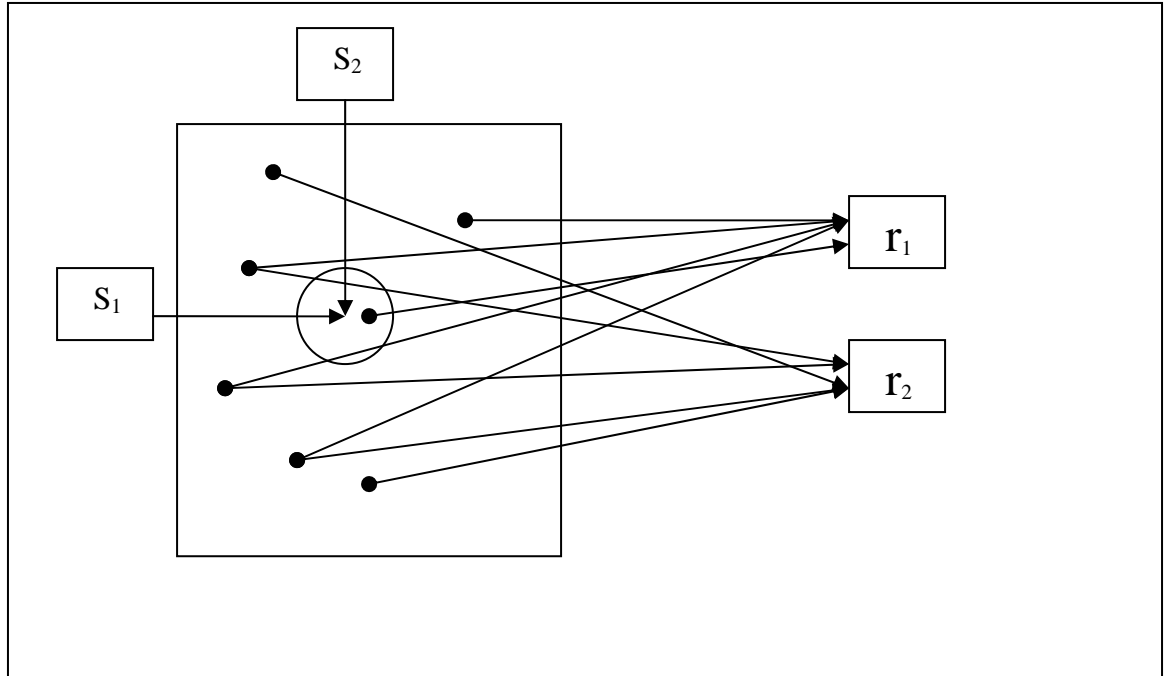
$$d' = \ln\left(\frac{h}{1-h}\right) - \ln\left(\frac{f}{1-f}\right) = (c \cdot r_{\text{old}} - \beta) - (c \cdot r_{\text{new}} - \beta) = c \cdot (r_{\text{old}} - r_{\text{new}}).$$

Thus  $d'$  directly measures the difference in activation by old and new stimuli.

In sum, the prototype model has five model parameters: the discriminability parameter  $\sigma$  which determines the width of the generalization gradients; the learning rate parameter  $\alpha$  for the delta learning rule; the sensitivity parameter  $b$  for the categorization choice rule; and two parameters for the recognition response rule – a sensitivity parameter  $c$  and a response bias parameter  $\beta$ .

*Exemplar Model.* The exemplar model is illustrated in Figure 4 shown below. This model is also based on three sets of assumptions. First, the inputs to the network are assumed to form a square grid, with each point on the grid representing a single input node. The stimulus  $S(t) = [s_1(t), s_2(t)]$  activates a circular receptive field of grid points. The centroid of the receptive field is located at the pair of stimulus values  $(s_1, s_2)$ . The amount of activation of a nearby input node declines as a function of the distance of the node from this center. Second, these inputs are passed through a set of weighted connections to the output nodes corresponding to each category, which are used to generate the category response. Finally, the connection weights are updated on the basis of feedback following each response during training. The details about the assumptions are presented next.

Figure 4: A Connectionist Version of an Exemplar Model



The exemplar model assumes that the stimulus is represented by a square grid (or square table) of input nodes, with  $m$  rows and  $m$  columns. Each node on the grid (or cell of the table) is designed to detect a pair of stimulus values. In particular, the node in the cell corresponding to row  $i$  and column  $j$  is designed to detect the value  $X_{ij} = [X_i, X_j]$ , which is called the ideal point for this node. The activation this node, denoted,  $x_{ij}(t)$ , is determined by the similarity of the stimulus,  $S(t)$  to the ideal point  $X_{ij}$ , denoted  $sim(X_{ij}, S)$ . A Gaussian type of generalization gradient is used to form the receptive field:

$$sim(X_{ij}, S) = e^{-\left(\frac{X_i - s_i}{\sigma}\right)^2} \cdot e^{-\left(\frac{X_j - s_j}{\sigma}\right)^2} \quad (6a)$$

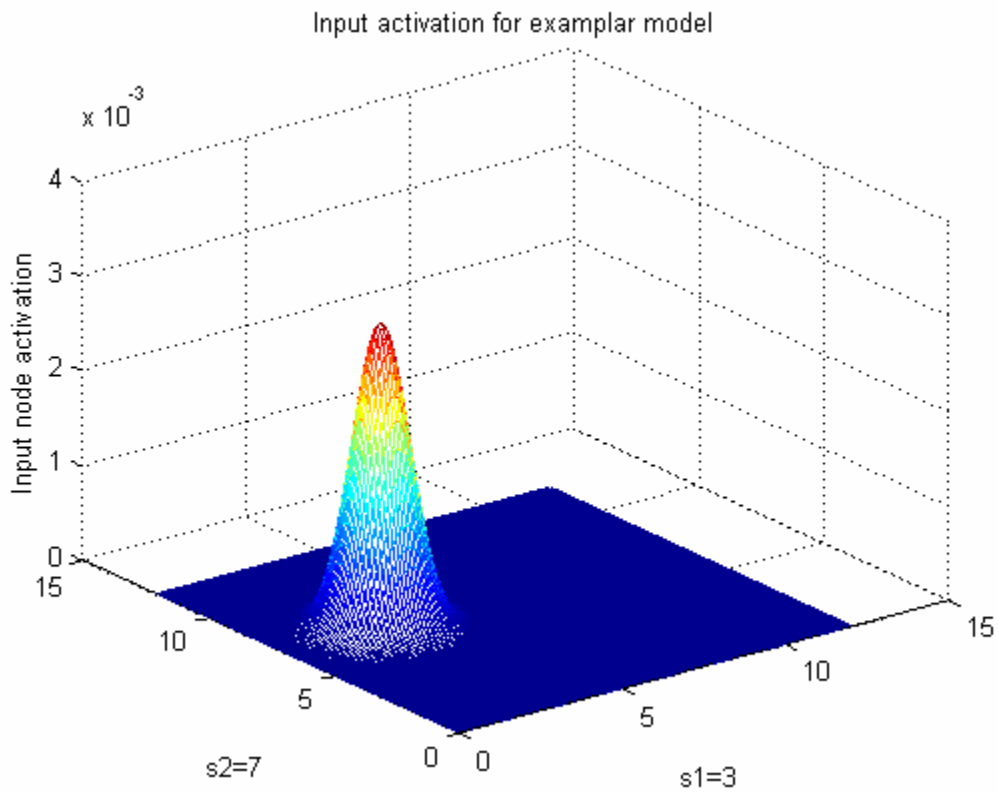
The parameter  $\sigma$  is the discriminability parameter which determines the width of the generalization gradient of the receptive fields. Low discriminability (large values of  $\sigma$ ) produce a large receptive field, which makes it hard to detect differences among stimuli. High discriminability (small values of  $\sigma$ ) produce a small receptive field, which makes it easy to detect differences among stimuli.

The activation corresponding to the node in row  $i$  and column  $j$  of the grid,  $x_{ij}(t)$ , is determined by the similarity for this node relative to the sum of all the similarities in the grid or table:

$$x_{ij}(t) = \frac{\text{sim}(X_{ij}, S)}{\sum \sum \text{sim}(X_{ij}, S)}. \quad (6b)$$

A stimulus produces a bivariate distribution of input activations on the grid, which is centered around the pair of stimulus values. Figure 5 illustrates the bivariate distribution of activation produced by the stimulus  $S(t) = [3,7]$ . To generate this figure, we used  $m = 121$  equally spaced nodes with ideal points covering the stimulus range from 0 to 12, and we set the discriminability parameter equal to  $\sigma = 1$ .

Figure 5: Input activations produced by the exemplar model.



The input nodes are connected to two category nodes, one for each category. The activation of the two category nodes are denoted  $r_A(t)$  and  $r_B(t)$  for category A and B, respectively. Each point on the grid or each cell of the table has a connection weight connecting an input node to a category. The connection weight,  $w_{ij,k}(t)$ , connects the input activation  $x_{ij}(t)$  to the k-th category output. The activation of the category nodes is based on the following linear input to output mapping:

$$r_k(t) = \sum_i \sum_j w_{ij,k}(t) \cdot x_{ij}(t) . \quad (7)$$

In other words, we multiply the connection weight in each cell by the input activation for that cell, and sum across all the cells in the table of input nodes.

This model is called an exemplar model because each receptive field of a training stimulus is associated with the output category nodes through a separate set of connection weights. Thus the model simply associates each region of the stimulus space with a response, and similar examples get mapped to similar responses.

The connection weights are updated according to an error reduction or delta learning rule:

$$w_{ij,k}(t+1) = w_{ij,k}(t) + \alpha \cdot [f_k(t) - r_k(t)] \cdot x_{ij}(t) \quad (8)$$

This is the same learning model that was used for the prototype model.

For a categorization task, the probability of choosing category A to a given stimulus is given by the ratio rule given earlier as Equation 4; and for a recognition task, the probability of responding old to a stimulus is determined by sum of activations as expressed in Equation 5.

In sum, the exemplar model, like the prototype model, has four crucial model parameters: the discriminability parameter  $\sigma$  which determines the width of the stimulus

generalization gradients; the learning rate parameter  $\alpha$  for the delta learning rule; the sensitivity parameter  $b$  for the categorization choice rule; and two parameters for the recognition response rule – a sensitivity parameter  $c$  and a response bias parameter  $k$ .

The two models share many assumptions including the use of Gaussian generalization gradients for the input nodes, linear mappings from inputs to outputs, and choice response rules. The main difference between the two models is in terms of the input representation: in the first case, two univariate sets of input nodes were used, whereas in the second case a single bivariate grid of input nodes was used.

### 3. Qualitative Comparison of Models.

Despite the similarity of the two models, they make qualitatively different predictions with regard to the special transfer stimuli. We can mathematically prove that the prototype model cannot predict the crossover interaction observed in Table 1, and we can show through computer simulation that the exemplar model does predict this interaction for a wide range of parameter values.

*Predictions of prototype model.* First we prove that the prototype model cannot produce the crossover interaction shown in Table 1. To do this, it will be useful to rewrite Equation 4 into a more convenient form:

$$\Pr[A | S(t)] = \frac{e^{b \cdot r_A(t)}}{e^{b \cdot r_A(t)} + e^{b \cdot r_B(t)}} \cdot \frac{e^{-b \cdot r_A(t)}}{e^{-b \cdot r_A(t)}} = \frac{1}{1 + e^{b[r_A(t) - r_B(t)]}} \quad (9)$$

This new expression makes it clear that the probability of choosing A is an increasing function of the difference

$$\begin{aligned} [r_A(t) - r_B(t)] &= \{\sum_i v_{iA} \cdot x_i(t) + \sum_i w_{iA} \cdot y_i(t)\} - \{\sum_i v_{iB} \cdot x_i(t) + \sum_i w_{iB} \cdot y_i(t)\} \\ &= \sum_i (v_{iA} - v_{iB}) \cdot x_i(t) + \sum_i (w_{iA} - w_{iB}) \cdot y_i(t) \end{aligned} \quad (10)$$

The time index has been dropped from the connection weights because the transfer tests in Table 1 occur after training, and no more feedback is provided. Thus we assume that the connection weights are fixed at this point. We note that  $x_i(t)$  is solely a function of  $s_1(t)$  and so we can rewrite the first sum in Equation 10 more conveniently as

$$V[s_1(t)] = \sum_i (v_{iA} - v_{iB}) \cdot x_i(t).$$

Similarly, we note that  $y_i(t)$  is solely a function of  $s_2(t)$  and so we can rewrite the second sum in Equation 10 more conveniently as

$$W[s_2(t)] = \sum_i (w_{iA} - w_{iB}) \cdot y_i(t).$$

Using this new notation, when stimulus  $S = (s_1, s_2)$  is presented for categorization, then

$$(r_A - r_B) = V[s_1] + W[s_2]. \quad (10)$$

In other words, the probability of choosing A is an increasing function of the additive effects of the first and second dimension values.

Fundamental measurement theorists (Krantz et al, 1972) call Equation 10 an additive conjoint measurement model, and this class of models must satisfy an ordinal property called the independence axiom. This axiom is derived below. We start by noting that the second row of Table 2 implies

$$\begin{aligned} \Pr[ A | S = (1,1) ] &< \Pr[ A | S = (1,10) ] \\ \rightarrow V[s_1 = 1] + W[s_2 = 1] &< V[s_1 = 1] + W[s_2 = 10] . \end{aligned}$$

Canceling the common term from the left and right hand side yields

$$\rightarrow W[s_2 = 1] < W[s_2 = 10] ,$$

and adding a common term to the left and right hand side yields

$$\begin{aligned} \rightarrow V[s_1 = 10] + W[s_2 = 1] &< V[s_1 = 10] + W[s_2 = 10] \\ \rightarrow \Pr[ A | S = (10,1) ] &< \Pr[ A | S = (10,10) ] . \end{aligned}$$

Thus we have proved that the rank order of the columns observed in the first row of Table 1 must be the same as the rank order of the columns in the second row of Table 1. This is the independence axiom implied by additive conjoint measurement models. The crossover interaction violates this independence axiom, ruling out the additive conjoint measurement model, and thus disconfirming the predictions of the prototype model.

It is important to note that this test of the prototype model is very robust with respect to a number of ad hoc assumptions made regarding the prototype model. For example, suppose we changed the generalization gradient from the Gaussian activation function shown in Equation 1a to another activation function such as an exponential gradient. Changing this assumption has no effect on this test, because the prototype model still must satisfy the independence axiom. (Changing the number of input nodes,  $m$ , for each set also has no effect on this test.) Suppose we change the choice probability rule in Equation 4 to any other monotonically increasing function of the difference ( $r_A - r_B$ ). Once again, changing this assumption has no effect on this test because it is only sensitive to the ordinal relations among the choice probabilities. Suppose we change the learning rule from the delta rule to some other learning rule. Again this has no effect, and the independence axiom still must be satisfied by the prototype model. Finally, we note that this test of the prototype model does not depend on any specific values of the model parameters, and instead, it holds true for all parameter values.

*Predictions of exemplar model.* Turning to the exemplar model, how robustly does this model predict the crossover interaction effect shown in Table 2? It is difficult to prove that the exemplar model must predict the crossover interaction effect. An alternative method of analysis for this type of situation is to use computer simulation to

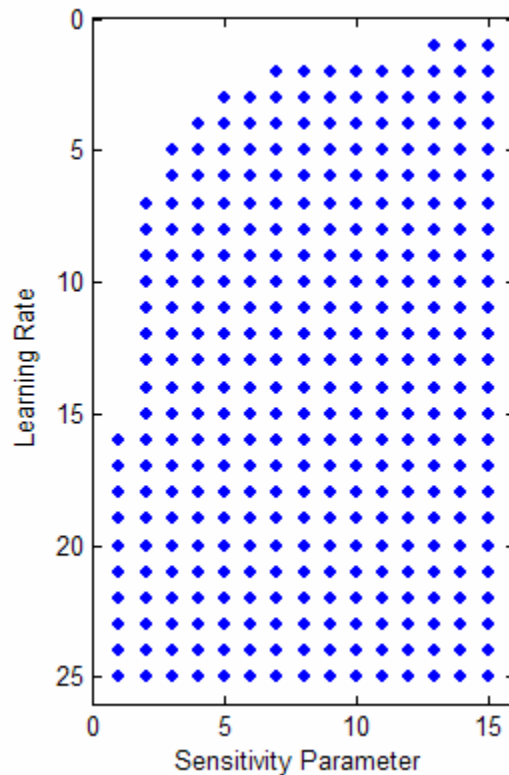
examine a feasible range of the parameter space. This involves (1) selecting a set of parameter values; (2) looping through Equations 6, 7, and 8 for the 400 training stimuli; (3) fixing the connection weights after training, and computing the predictions for the 4 special transfer test stimuli using Equations 6, 7 and 4; (5) evaluate whether or not the predictions successfully reproduce the crossover interaction pattern; (6) repeat this process with a new selection of parameter values until a wide range of values in the parameter space has been examined. This way we can examine the extent to which the model supports a particular prediction across the entire parameter space.

For the categorization task, the exemplar model has three critical parameters: one is the discriminability parameter,  $\sigma$ , which determines width of the generalization gradient for the receptive fields; a second is the learning rate,  $\alpha$ , for the delta learning rule; and a third is the sensitivity parameter  $b$ , for choice probability. A computer simulation of the model was conducted with  $\sigma$  ranging from 1 to 10 in unit steps,  $\alpha$  ranging from .04 to 1.0 in .04 increments; and  $b$  ranging from 1 to 15 in unit steps. This generated  $10 \times 25 \times 15 = 3750$  simulations. For each simulation, we checked whether or not the model predicted a difference between the column proportions in the correct direction, separately for each row. Also, the difference had to exceed a cutoff equal to .20 in magnitude before it was counted as a success. (The appendix describes the actual program used to compute these results).

The results of these simulations can be summarized as follows. The model either reproduces the correct crossover interaction pattern, or it predicts a difference that is too small to detect. Thus it never predicted a pattern that satisfies the independence axiom for the parameter values that we examined. Whenever the discriminability parameter exceeds

$\sigma > 5$ , the correct pattern is successfully reproduced; after that the correct pattern occurs whenever the learning rate parameter  $\alpha$  and sensitivity parameter  $b$  are sufficiently large. In all other cases, the correct pattern is predicted but the difference is below the cutoff. This is illustrated in Figure 4 shown below, which was generated by the exemplar model with  $\sigma = 5$ , producing 375 test points. The horizontal axis represents the 15 values of the sensitivity parameter,  $b$ , increasing from left to right; and the vertical axis represents the 25 values of the learning rate parameter  $\alpha$  increasing from top to bottom. Each filled point indicates a combination of parameters that successfully reproduced the crossover interaction; and the unfilled points indicate parameters that failed to produce a detectable difference. The total number of combinations that reproduced the crossover turns out to be 337 out of 375 possibilities for  $\sigma = 5$ . As seen in this figure, the model successfully reproduces the crossover whenever all three parameters are sufficiently large.

Figure 6: Points in the parameter space where the exemplar model predicts the correct pattern of results for the four special transfer stimuli of Table 1.

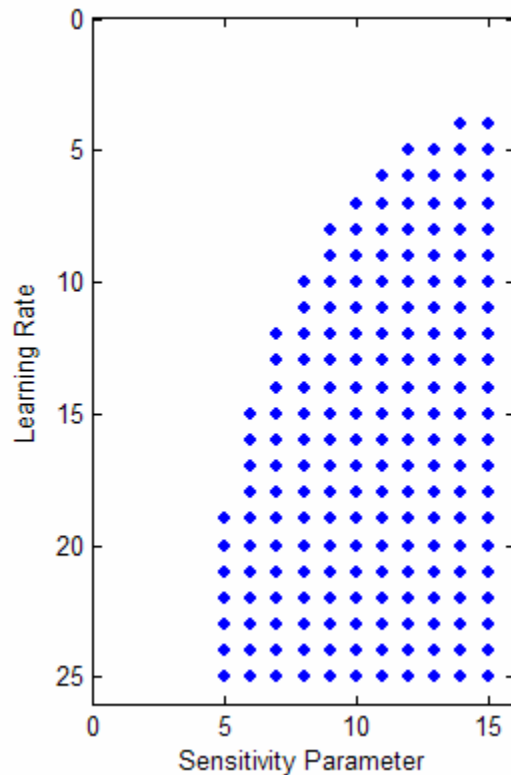


As a check on our theory, we also ran the same computer simulation using the prototype model to generate the predictions, rather than the exemplar model. According to our theoretical test of independence, none of the parameters should be able to reproduce the interaction shown in Table 1. In fact, this is what we found when we checked the same 3750 combination of parameters described above. Although computer simulation is never a substitute for a mathematical proof, it is useful to check the proof with computer simulation. Sometimes there is a hidden assumption in a proof that is not imposed on the computer simulation, and so the two methods could produce different results.

*Experimental Design.* The importance of using a well controlled experiment with carefully designed transfer stimuli for comparing the models needs to be emphasized. The independence property that we established for the prototype model depends on

manipulating one stimulus dimension while holding the other dimension constant in Table 1. Suppose we were not so careful with the design of the four special transfer stimuli located at the center of each cluster in Figure 1. Suppose that the four clusters are slightly perturbed and they are centered at positions [1,1], [2,10], [9,1], and [10,9] instead. Furthermore, suppose we observed a similar pattern of results as shown in Table 1 for these four new conditions. In this case, we cannot perform the test of independence – these stimuli fail to satisfy the criteria of holding one dimension constant while manipulating the other. With this design, we would have to resort to computer simulation to test whether or not the prototype model predicts the pattern of results, just as we did the exemplar model. In fact, if we compute the predictions from the prototype model using the same 375 combination of parameters that we examined with the exemplar model (with  $\sigma = 5$ ), then we find the results shown in Figure 7 below. This figure shows that over 70% of the parameters succeed in reproducing the correct pattern. In sum, without the proper experimental design, it is very difficult to discriminate the prototype from the exemplar model.

Figure 7: Points in the parameter space where the prototype model predicts the correct pattern when the four transfer stimuli in Table 1 are not properly controlled.



#### 4. One or two memory systems?

Next we turn to a qualitative analysis of a theoretical issue raised by the findings shown in Table 2. Recall that the amnesic participants performed slightly better than the normal controls on the categorization task; but the amnesic participants performed much worse than the normal controls on the recognition task. It seems that the amnesic participants could categorize the stimuli without recognizing them. These results suggest that an explicit memory system, used for recognition, is damaged in the amnesic participants; while the implicit memory system, used for categorization, remains intact. This raises an important theoretical question: Can these results be explained by a single memory system (e.g. the exemplar model), or must resort to a dual memory model to account for these results?

The first step in this analysis is to start out by examining the predictions of a single exemplar based memory system for both the categorization and recognition performance. To account for differences between the amnesic and normal individuals, we assume that some model parameters differ across these two types of populations. At this point, we need to state some plausible assumptions that relate the parameters of the exemplar model to the individual differences between the normal and amnesic populations.

The hypothesis that we will examine is that the discriminability parameter differs across amnesic and normal individuals (see Nosofsky & Zaka, 1998, for the original version of this hypothesis). The discriminability parameter is a plausible candidate because it reflects the distinctiveness of the stimulus – response associations. A low discriminability parameter produces a diffuse set of associations, making it difficult to discriminate representations in memory; whereas a high discriminability parameter produces highly distinctive associations, making it easy to discriminate representations in memory. In sum, we assume that a single memory system operates, that is the exemplar model, but the normal participants have a higher discriminability parameter (producing more distinctive memories) as compared to the amnesic individuals. So the question now becomes – can this simple change in this single parameter across the two types of individuals explain the interaction shown in Table 2?

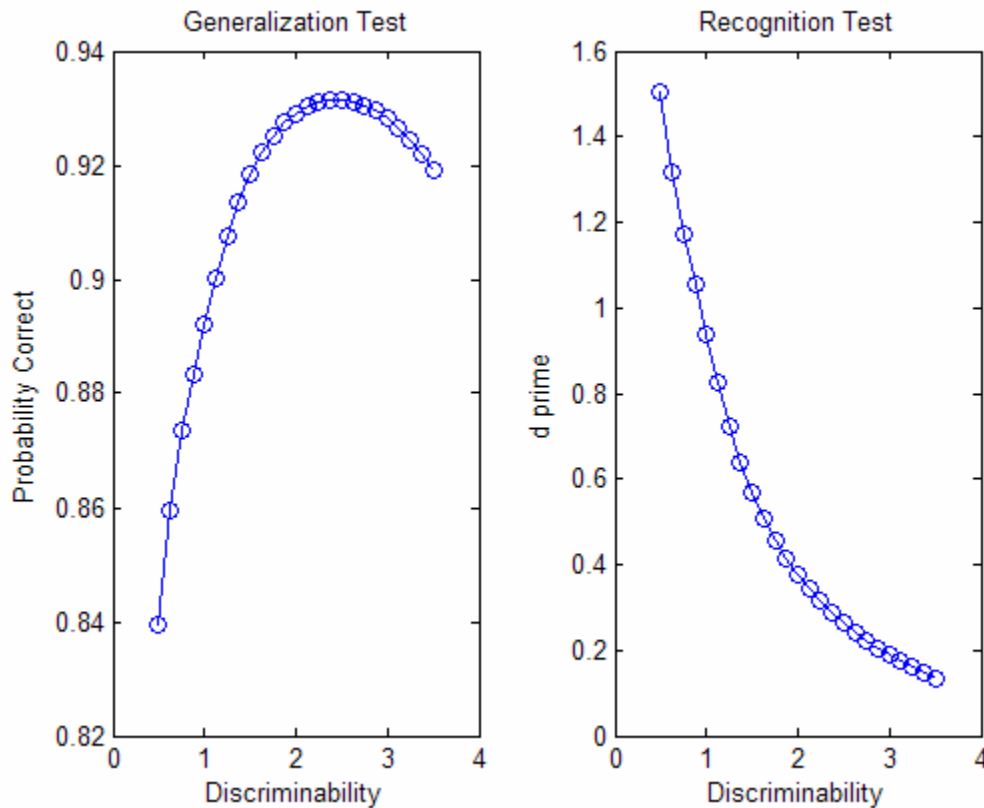
To evaluate this single memory explanation, we can use computer simulations to examine the model predictions for categorization and recognition across a wide range of model parameters. For this analysis, the learning rate and the response sensitivity parameters were not critical to the results, and so they were simply fixed to the same

values for both the normals and the amnesics for all of the simulations. The critical discriminability parameter,  $\sigma$ , was varied from .5 to 3.5 in .125 increments.

The results of these simulations are shown in Figure 8 below. The left panel shows the predicted results for categorization of the generalization test stimuli, and the right panel shows the predicted results for recognition of old training from new transfer stimuli. The horizontal axis on each panel represents the values of the discriminability parameter, and the vertical axis represents performance on each task. Recall that small values of  $\sigma$  produce high levels of discriminability, and large values of  $\sigma$  produce low levels of discriminability. The predictions turn out to be highly nonlinear and very counterintuitive. On the one hand, categorization performance is a non-monotonic function of discriminability – it increases and then decreases as a function of  $\sigma$ . On the other hand, recognition performance is a monotonically decreasing function of  $\sigma$ .

What are the implications of these predictions for a single versus dual memory system explanation of the findings in Table 2? Suppose that normal subjects tend to have a high level of discriminability (e.g.,  $\sigma = 1$ ) whereas the amnesics tend to have a low level of discriminability (e.g.,  $\sigma = 3$ ). Then the single memory model successfully reproduces the puzzling pattern of findings in Table 2. If we choose these values of discriminability for the normal and amnesic groups, then the amnesic group performs slightly better than the normal group on categorization, while at the same time, the normal group greatly exceeds the amnesic group in terms of recognition performance. In fact, recognition performance for the amnesic group's is predicted to be close to zero, even though categorization performance for this same group is predicted to be close to perfect.

Figure 8: Predictions of the Exemplar Model for Categorization and Recognition



In summary, an intuitively appealing explanation for the puzzling results of Table 2 is that a dual memory system is operating. However, a more rigorous cognitive modeling analysis indicates that this conclusion is premature. The data can be easily explained by a single memory system. This example shows the importance of examining explanations at a more rigorous level of analysis, rather than relying on intuitive reasoning. The above analysis was inspired by Nosofsky and Zaka's (1998) re-analysis of the Kowlton and Squires (1992) data. Nosofsky and Zaka (1998) demonstrated that a single memory system model, in particular, an exemplar model, could account for both the categorization and recognition performance of the amnesic and normal group data

reported by Knowton and Squire. This is a very good example of the use of cognitive models for evaluating basic theoretical issues.

references

Knowlton, B. J. & Squire, L. (1993) The learning of categories: parallel brain systems for item memory and category knowledge. *Science*, 262, 1747-1749.

Nosofsky, R. M. & Zaka, S. F. (1998) Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar based interpretation. *Psychological Science*, 9, 247-255.