Dimensional Attention Learning in Models of Human Categorization

John K. Kruschke
Department of Psychology and Cognitive Science Program
Indiana University, Bloomington IN 47405 USA
e-mail: kruschke@ucs.indiana.edu

Abstract

When humans learn to categorize multidimensional stimuli, they learn which stimulus dimensions are relevant or irrelevant for distinguishing the categories. Results of a category learning experiment are presented, which show that categories defined by a single dimension are much easier to learn than categories defined by the combination of two dimensions. Three models are fit to the data, ALCOVE (Kruschke 1990a,b, in press), standard back propagation (Rumelhart, Hinton & Williams 1986), and the configural-cue model (Gluck & Bower 1988). It is found that ALCOVE, with its dimensional attention learning mechanism, can capture the trends in the data, whereas back propagation and the configural-cue model cannot. Implications for other models of human category learning are discussed.

Introduction

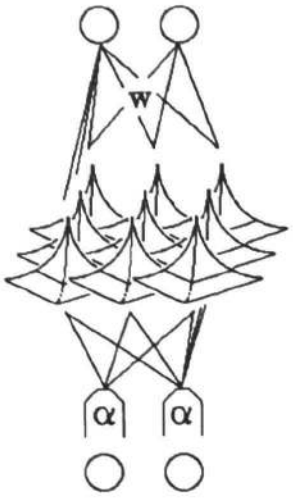
Imagine learning to classify mushrooms as "poisonous" or "edible." You are shown one mushroom after another and told which category it belongs to. After seeing many examples, your accuracy of classifying new examples improves. One of the key aspects of such learning is the determination of which features of the mushrooms are relevant to the categorization. For example, it might be that all red mushrooms are poisonous, but red mushrooms occur in the same range of sizes as edible mushrooms. In that case it would be wise to weigh information about about color more heavily than information about size. On the other hand, it might be that more than one dimension is relevant to the categorization; e.g., perhaps red or spotted mushrooms are poisonous. In that case, one should pay attention to both dimensions of color and texture.

Posner (1964) called situations in which there was a single relevant dimension *gating* tasks, since the irrelevant dimension(s) could be gated out of consideration, and he called situations in which more than one dimension was relevant *condensation* tasks, since the information from multiple dimensions had to be condensed into a single categorization decision. Posner and others (e.g., Garner 1974) have established that gating tasks are generally easier to learn than condensation tasks. In this article I show that standard back-propagation (Rumelhart, Hinton & Williams 1986) and the configural-cue model (Gluck & Bower 1988) cannot capture that basic result, while another connectionist model called ALCOVE (Kruschke 1990a,b, in press) can. I report results of a category learning experiment and fits of the models to the data. The key difference between the models is that ALCOVE incorporates constraints to reflect the dimensional attention learning abilities of people, whereas the other models do not.

The ALCOVE Model

ALCOVE is a feed-forward network that learns by gradient descent on error, but it is unlike standard back propagation (Rumelhart *et al.* 1986) in its architecture, its behavior, and its goals. Unlike the standard back-propagation network, which was motivated by generalizing neuron-like perceptrons, the architecture of ALCOVE was motivated by a molar-level psychological theory, Nosofsky's (1986) generalized context model (GCM). The psychologically constrained architecture results in behavior that captures the detailed course of human category learning in many situations where standard back propagation fares less well (Kruschke 1990a,b, in press). And, unlike many applications of standard back propagation, the goal of ALCOVE is not to discover new (hidden-layer) representations after lengthy training, but rather to model the course of learning itself, by determining which dimensions of the given representation are most relevant to the task, and how strongly to associate exemplars with categories.

Like the GCM, ALCOVE assumes that input patterns can be represented as points in a multidimensional psychological space, as determined by multi-dimensional scaling algorithms (e.g., Kruskal 1964; Shepard 1962). Each input node encodes a single psychological dimension, with the activation of the node indicating the value of the stimulus on that dimension. Figure 1 shows the architecture of ALCOVE,



Category nodes.

Learned association weights.

Exemplar nodes.

Learned attention strengths.

Stimulus dimension nodes.

Figure 1: The structure of ALCOVE. The pyramids in the hidden layer indicate the activation profile of hidden nodes, as determined by Equation 1, with $r = q = 1$.

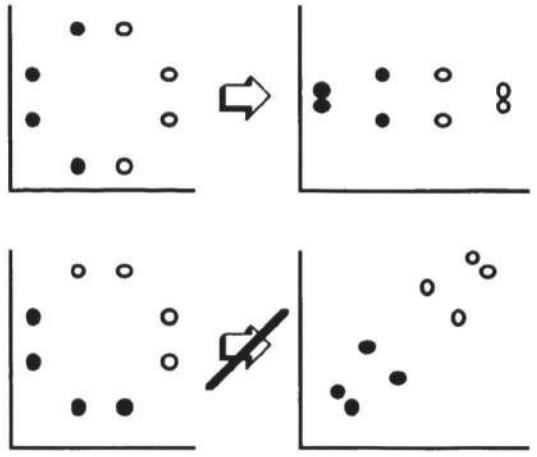


Figure 2: Top panel shows that increasing attention to the horizontal dimension and decreasing attention to the vertical dimension causes exemplars of the two categories (denoted by filled and open circles) to have greater between-category dissimilarity and greater within-category similarity (after Nosofsky 1986, Fig. 2). Lower panel shows that ALCOVE cannot differentially attend to diagonal axes.

illustrating the case of just two input dimensions.

Each input node is gated by a dimensional *attention strength* α_i. The attention strength on a dimension reflects the relevance of that dimension for the particular categorization task at hand, and the model learns to allocate more attention to relevant dimensions and less to irrelevant dimensions.

Each hidden node corresponds to a position in the multi-dimensional stimulus space, with one hidden node placed at the position of every training exemplar. Each hidden node is activated according to the psychological similarity of the stimulus to the exemplar represented by the hidden node. The similarity function comes from the GCM and the work of Shepard (1962; 1987): Let the position of the j^{th} hidden node be denoted as $(h_{j1}, h_{j2}, \ldots)$, and let the activation of the j^{th} hidden node be denoted as a_j^{hid}. Then

$$a_j^{hid} = \exp\left(-c \left(\sum_i \alpha_i \left| h_{ji} - a_i^{in} \right|^r \right)^{q/r} \right) \qquad (1)$$

where c is a positive constant called the *specificity* of the node, where the sum is taken over all input dimensions, and where r and q are constants determining the similarity metric and similarity gradient, respectively. For separable psychological dimensions, the city-block metric ($r = 1$) is used, while integral dimensions might call for a Euclidean metric ($r = 2$; *cf.* Garner 1974, Shepard 1964). A city-block distance metric ($r = 1$) with exponential similarity gradient ($q = 1$) is used here (Shepard 1987).

The dimensional attention strengths adjust themselves so that exemplars from different categories become less similar, and exemplars within categories become more similar. Consider a simple case of eight stimuli that form the corners of an octagon in input space, as in Figure 2. The stimuli are mapped to one of two categories, as indicated by filled or open circles. When only one dimension is relevant, as in the top panel of Figure 2, ALCOVE learns to increase the attention strength on the relevant dimension, and to decrease the attention strength on the irrelevant dimension. By contrast, ALCOVE cannot stretch or shrink diagonally, as suggested in the lower panel of Figure 2. Two points made in this article are (1) to demonstrate that this constraint is an accurate reflection of human performance, in that categories separated by a diagonal boundary take longer to learn than categories separated by a boundary orthogonal to one dimension, and (2) to show that standard back propagation and configural-cue model do not capture this fact.

Each hidden node in ALCOVE is connected to output nodes that correspond to response categories. The connection from the j^{th} hidden node to the k^{th} category node has a connection weight denoted w_{kj}, called the *association weight* between the exemplar and the category. The output (category) nodes are activated by the linear rule used in the GCM and the network models of Gluck and Bower (1988):

$$a_k^{out} = \sum_{\substack{hid \\ j}} w_{kj} a_j^{hid}. \qquad (2)$$

In ALCOVE, unlike the GCM, the association weights

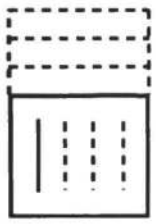


Figure 3: Stimuli used in category learning experiment. The rectangle could have one of four heights (the shortest rectangle is shown with a solid line) and the interior vertical segment could have one of four lateral positions (the leftmost position is shown with a solid line).

are learned and can take on any real value, including negative values. Category activations are mapped to response probabilities using the same choice rule (Luce 1963) as was used in the GCM and network models. Thus,

$$\Pr(K) = \exp(\phi\, a_K^{out}) \Big/ \sum_{\substack{out \\ k}} \exp(\phi\, a_k^{out}) \qquad (3)$$

where ϕ is a real-valued scaling constant. In other words, the probability of classifying the given stimulus into category K is determined by the magnitude of category K's activation relative to the sum of all category activations.

The dimensional attention strengths, α_i, and the association weights, w_{kj}, are learned by gradient descent on sum-squared error, as used in standard back propagation (Rumelhart *et al.* 1986) and in the network models of Gluck and Bower (1988). Space constraints prohibit further discussion of the model; details can be found in Kruschke (1990a,b, in press).

In fitting ALCOVE to human learning data, there are four free parameters: the fixed specificity c in Equation 1; the probability mapping constant ϕ in Equation 3; the association weight learning rate; and, the attention strength learning rate.

A Human Learning Experiment

To test the implications of attention learning in ALCOVE, human subjects were trained on the category structures shown in Figure 2. The stimuli were geometric forms, as shown in Figure 3. Each stimulus consisted of a rectangle with one of four heights, and an interior vertical segment at one of four lateral positions. Only 8 of the 16 possible combinations of height and position were used, corresponding to the abstract structure in Figure 2.

Scaling the stimulus space

The first step in modelling this situation is to determine the psychological coordinates of the stimuli. To do this, one obtains similarity ratings of pairs of stimuli, and determines the coordinate values in psychological space that best predict those similarities (Kruskal 1964; Shepard 1962). This process is analogous to generating a spatial map of cities when all you are told is the distances (dis-similarities) between cities.

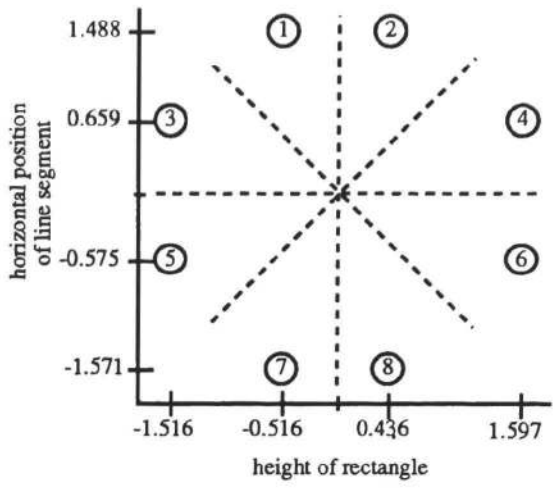


Figure 4: Locations of stimuli in psychological similarity space. Dashed lines suggest the four category boundaries used in the learning experiment.

Procedure: The two stimuli of each pair were presented sequentially on a computer screen for 1.75 seconds each, separated by a 0.75 second blank screen. The subject then entered a similarity rating from 1 to 9 on the computer keyboard. Each pair was presented four times in each order, for each of 50 subjects, yielding 400 similarity ratings for each pair.

The best-fitting psychological coordinates of the stimuli accounted for over 98% of the variance in similarity ratings, and are shown in Figure 4. Notice that while the *physical* values of height and horizontal position were equally spaced (see Figure 3), the *psychological* values were not. Note also that the two dimensions are about equally salient, in that the total range is about the same on the two dimensions, but the middle interval on the horizontal-position dimension is bigger than the middle interval of the height dimension. That implies that a category boundary indicated by the horizontal dashed line in Figure 4 should be easier to learn than a category boundary indicated by the vertical dashed line. This will be confirmed in the learning data.

Category learning results

The strong prediction of ALCOVE is that the four alternative category disctinctions indicated in Figure 4 should not be equally easy; rather, the distinctions for which only a single dimension is relevant (the vertical

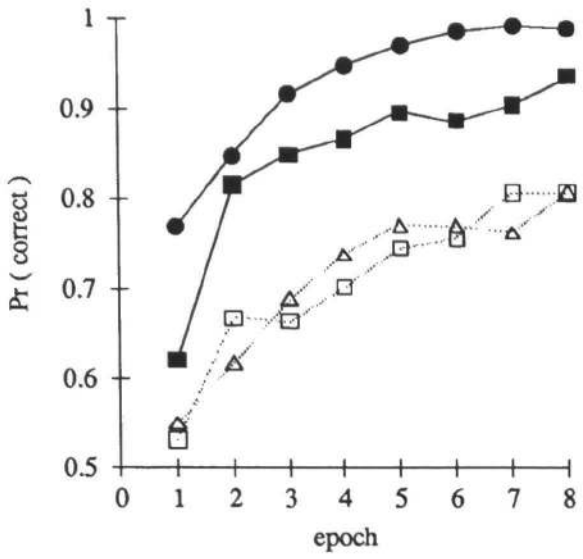


Figure 5: Human learning data. Filled markers correspond to single-dimension categories (filled circle is horizontal position relevant, filled square is height relevant); open markers show diagonal categories.

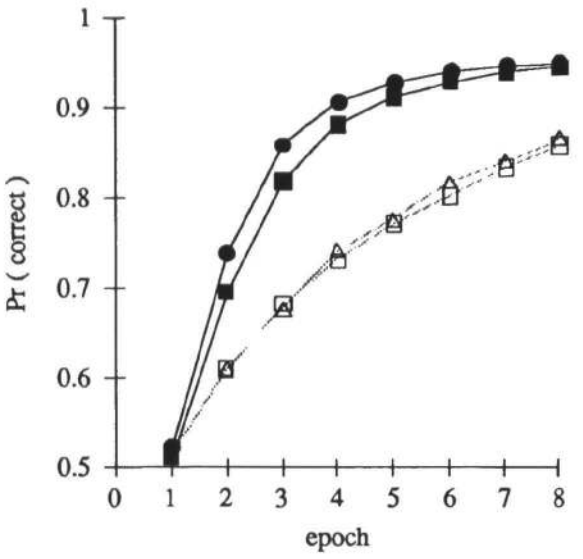


Figure 6: Best fit of ALCOVE to human learning data in Figure 5.

and horizontal dashed lines in Figure 4) should be easier than the diagonal distinctions, because attention learning cannot differentially accentuate diagonal directions. The four category distinctions were given to different groups of subjects. For example, one group learned the position-relevant distinction, for which the stimuli marked 1, 2, 3 and 4 in Figure 4 were given one category label, with the remaining stimuli (numbered 5–8) were given a different category label.

Procedure: Subjects were given instructions that included exposure to the eight stimuli without any category feedback. Each training trial consisted of a presentation of a stimulus, which was terminated when the subject pressed a response key. The subject was then given feedback indicating whether the response was correct or incorrect, and the correct response. Each of the four groups saw the same fixed sequence of 64 stimuli. All that varied between groups was the category labels given to the stimuli. A total of 160 subjects were run, 40 in each group. Category labels were counter-balanced within groups.

Results are summarized in Figure 5. Each datum shows the mean percent correct for the preceding 8 trials (one epoch). Two effects are clear: The single-dimension categories are learned much faster than the diagonal categories; and, the horizontal-position dimension is learned faster than the height dimension.

Modelling the learning

Fit of ALCOVE

ALCOVE was applied to this situation by using two input nodes, corresponding to the two stimulus dimensions, eight hidden nodes, corresponding to the eight training exemplars, and two output nodes, corresponding to the two categories. The task for the model was to fit the learning curves from the four category types using a single set of parameter values. The discrepancy of ALCOVE from the human data was measured as the squared difference between observed category choice probabilities and predicted choice probabilities, summed across the 64 individual trials in each category type, across both category choices within each type, and across the four types.

The best fit of ALCOVE is shown in Figure 6. The fit produced a root mean squared deviation (RMSD) of 0.116, with parameter values of $\phi = 1.568$, $c = 1.662$, association weight learning rate of 0.08431 and attention learning rate of 0.6593. ALCOVE clearly shows the two main trends seen in the human data: The single-dimension categories are learned much faster than the diagonal categories, and the horizontal-position dimension is learned faster than the height dimension. ALCOVE learns the single-dimension categories faster by virtue of attention learning — the attention strength on the relevant dimension increases rapidly, while the attention strength on the irrelevant dimension decreases rapidly. ALCOVE learns the position dimension faster than the height dimension because the exemplar nodes reflect the greater distinctiveness of the middle interval of the position dimension.

Fit of back propagation

Standard back propagation (henceforth "backprop") was also applied. Two input nodes and two output nodes were used, as in the application of ALCOVE. In order to equilibrate the backprop architecture with the

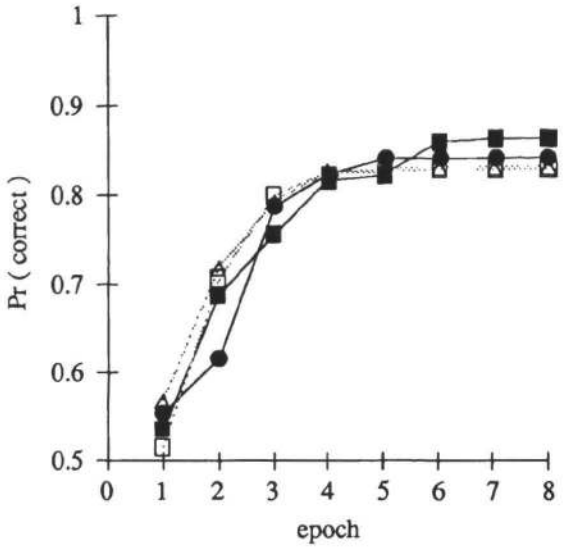


Figure 7: Best fit of standard back propagation to human learning data in Figure 5.

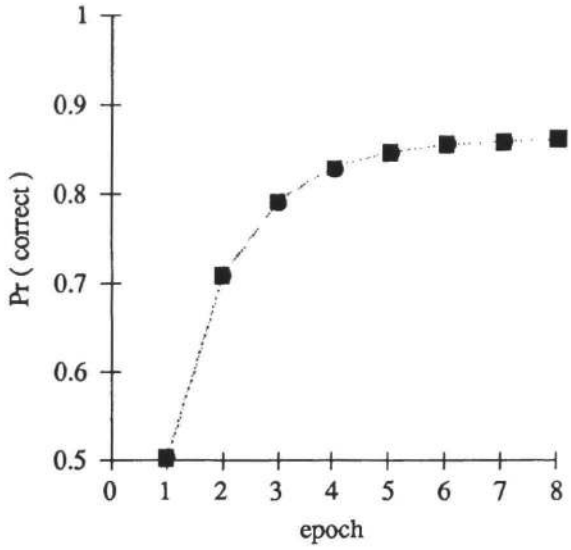


Figure 8: Best fit of the configural-cue model to human learning data of Figure 5. All four curves are exactly superimposed.

ALCOVE architecture as much as possible, the two output nodes were linear, with response probabilities computed using Equation 3, and eight hidden nodes were used (with standard linear-sigmoid activation functions; Rumelhart *et al.* 1986).

The backprop network was given five free parameters: The learning rate on the output weights; the learning rate on the hidden weights; the learning rate on the hidden node thresholds; the response mapping constant ϕ in Equation 3; and, the value of a fixed *gain* parameter in the sigmoidal activation function (see Kruschke & Movellan 1991). Initial values of hidden weights and thresholds were drawn from a uniform distribution over the interval $[-1, +1]$.

For any choice of parameter values, the fit was measured in the same way as for ALCOVE, but choice predictions of backprop were computed by first averaging over 200 different random initializations of the hidden weights and thresholds.

The best fit of backprop to the learning data is shown in Figure 7. It yielded an RMSD of 0.152, using $\phi = 0.6636$, output weight learning rate of 0.2049, a hidden weight learning rate of 1.091, a hidden threshold learning rate of 0.04159, and a gain of 2.249. The fit of backprop is much worse than that of ALCOVE, and the qualitative behavior of backprop departs badly from the data. Indeed, backprop learns the single-dimension and diagonal categories at essentially the same pace. The best backprop can do is try to match the mean learning curve across all four category types.

Why does backprop do so poorly? To explain why, first I'll define a hidden node's *weight vector* as the ordered list of connection weights fanning into the node. The weight vector of a hidden node specifies the direction in stimulus space that causes the biggest change in the node's activation value, and thereby indicates the underlying dimension in stimulus space to which the node is responding. The weight vectors can, potentially, point in *any* direction in stimulus space. In particular, they can align with the diagonal axes of the stimulus space just as easily as they can align with the canonical axes. Standard back-propagation learning is *isotropic* in that sense, unlike human learning.

Fit of the configural-cue model

The configural-cue model (Gluck & Bower 1988) was originally proposed for stimuli with binary-valued dimensions, and therefore is not directly applicable to the present situation. However, it can be reasonably extended as follows: Each value on each dimension is encoded by a separate input node, and each combination of values from the two dimensions is encoded by a separate node, yielding $4 + 4 + 16 = 24$ input nodes. A given stimulus activates a subset of 3 of the 24 input nodes. The input nodes are connected directly to the output nodes, which are governed by Equations 2 and 3. The configural-cue model has two parameters, the learning rate for the connection weights and the mapping constant in Equation 3.

The best fit is shown in Figure 8 ($\phi = 0.9289$, connection weight learning rate was 0.1591, yielding an RMSD of 0.150). Figure 8 shows that all four learning curves are exactly superimposed, quite unlike the human data. The reason is that the (extended) configural-cue model makes no structural distinction between the category types in its input representation; the model has no representation of the fact that some

values lie on the same dimension but other values come from different dimensions.

Discussion

It is possible that back propagation or the configural-cue model could be modified to include some form of dimensional attention learning; this awaits future research. The results reported here pose a challenge for other models of category learning that do *not* incorporate attention learning, such as Hanson and Gluck's (1991) Cauchy-node model. It remains to be seen if such models can fit the data reported here, without extending them to include some form of dimensional attention learning. Other models that *do* include attention learning, such as J. R. Anderson's (in press) rational model, or Hurwitz's (1990) hidden pattern unit model, could, no doubt, capture the difference between single-dimension and diagonal categories.

All of these models, however, would probably fail to match one other prominent aspect of the data: the extreme rapidity with which the single-dimension categories are learned, very early in training (see Figure 5). Some subjects, in fact, made virtually no errors after the first two or three trials. There is little hope that learning algorithms that take small incremental steps on each trial could accomplish that.

What is needed, I believe, is a *rule hypothesizing system*: Early in training the subject might guess the right rule and make no more errors. Perhaps the primary question for such a rule generating system is, Which rules should be hypothesized and tested first? The behavior of ALCOVE suggests that one might generate and test rules using the dimension(s) with the largest attention strength. The notion is that category learners always employ an ALCOVE-like system, and simultaneously try to summarize, generalize, and leap-frog the performance of that system by hypothesizing and testing rules. The underlying ALCOVE-like system steers the rule generating system and acts as a fall-back when adequate rules are not yet found.

Acknowledgments

I thank Terry Bleizeffer, Steve McKinley, and Rita Randolph for running subjects. This research was supported by Biomedical Research Support Grant RR 7031-25 from the National Institutes of Health.

References

Anderson, J. R. (in press). The adaptive nature of human categorization. *Psychological Review.*

Garner, W. R. (1974). *The Processing of Information and Structure.* Hillsdale, NJ: Erlbaum.

Gluck, M. A. & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *J. of Memory and Language*, **27**, 166-195.

Hanson, S. J. & Gluck, M. A. (1991). Spherical units as dynamic consequential regions: Implications for attention, competition and categorization. In: R. P. Lippmann, J. Moody and D. S. Touretzky (eds.), *Advances in Neural Information Processing Systems 3.* San Mateo, CA: Morgan Kaufmann.

Hurwitz, J. B. (1990). A hidden-pattern unit network model of category learning. PhD dissertation, Harvard University.

Kruschke, J. K. (1990a). A connectionist model of category learning. PhD dissertation, University of California at Berkeley. Available from University Microfilms International.

Kruschke, J. K. (1990b). ALCOVE: A connectionist model of category learning. Research Report 19, Cognitive Science Program, Indiana University.

Kruschke, J. K. (in press). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review.*

Kruschke, J. K., & Movellan, J. R. (1991). Benefits of gain: Speeded learning and minimal hidden layers in back-propagation networks. *IEEE Trans. Systems, Man and Cybernetics*, **21**, 273-280.

Kruskal, (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115-129.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, **70**, 61-79.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *J. Exp. Psych.: General*, **115**, 39-57.

Posner, M. I. (1964). Information reduction in the analysis of sequential tasks. *Psychological Review*, **71**, 491-504.

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.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I & II. *Psychometrika*, **27**, 125-140, 219-246.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *J. of Mathematical Psychology*, **1**, 54-87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.