CHAPTER ELEVEN

Learning Functional Relations Based on Experience With Input–Output Pairs by Humans and Artificial Neural Networks

Jerome R. Busemeyer, Eunhee Byun, Edward L. Delosh and Mark A. McDaniel

I. DECISIONS, PREDICTIONS, AND ABSTRACT CONCEPTS

Before making any serious decision, we normally try to anticipate how the effects of our action will vary depending on the action taken. For example, before an anaesthetist can decide the amount of anaesthetic to administer to a patient, she needs to predict how the analgesic effect will vary as a function of the amount injected. Before a father can decide the amount of money to invest in his son's college education, he needs to predict how the return will vary as a function of the size of the investment. The point is that prediction is essential to decision making.

Predictions are thought to be based on knowledge of the *functional relation* between the strength of a cause and the magnitude of an effect. For this reason, there is a large body of empirical research by decision scientists investigating how people learn functional relations (Slovic & Lichtenstein 1971, Klayman 1988). Much of this research, however, has been not been synthesized and integrated into coherent theory, and so this literature remains disconnected from mainstream cognitive psychology.

From a cognitive perspective, functions can be viewed as abstract concepts that summarize cause–effect relationships. Cognitive psychologists have made great progress developing theories of how people learn abstract concepts (see Estes 1994). However, most of this theoretical effort has been restricted to one simple type of concept learning task called *categorization*. It is unclear whether or not theories of category learning can be extended for application to function learning.

The purpose of this chapter is to develop a concept learning model that can account for results from both categorization and function learning tasks. The remainder of the chapter is organized as follows: Section II discusses similarities and differences between category- and function-learning tasks, Section III synthesizes some basic findings on function-learning, Section IV describes an artificial neural network model of category learning and extends this model to function learning, and Section V shows how the extended model reproduces the basic findings from function learning.

II. CATEGORY VERSUS FUNCTION LEARNING PARADIGMS

There is considerable overlap in the basic experimental paradigms used to investigate category and function learning. In both cases, subjects are presented several hundred training trials, each of which consists of (a) the presentation of a stimulus called the *cue* (denoted $x(t)$ on trial t), (b) a response by the subject called the *prediction* (denoted $y(t)$ on trial t) and (c) feedback indicating the correct response called the *criterion* (denoted $z(t)$ on trial t).

For example, Koh & Meyer (1991) trained subjects to learn how to map a tone duration (*cue*) into a movement magnitude (*criterion*). On each trial, a tone duration was presented, the subject made a motor movement, and then the subject was shown the correct motor movement. This example involves mapping one physical continuum (x = tone duration) into a different physical continuum (z = movement magnitude). It is not necessary to employ different physical continua for stimuli and responses. For example, Delosh et al. (1996) trained subjects to map one line length cue into another line length criterion, thus employing a common physical continuum for stimuli and responses. (See Figure 11.1). Other researchers (e.g. Naylor & Clark 1968) used numbers to display the cue and criterion magnitudes.

After subjects learn the cue-criterion mapping for a set of training pairs, they are tested during a transfer phase on new stimuli never seen during training. The transfer test ascertains whether or not subjects can use the newly-learned concepts to interpolate or extrapolate.

For both category and function learning tasks, the mapping from cues to criteria may be probabilistic. In category learning, for example, disease A could occur on 60 per cent of the trials and disease B could occur on 40 per cent of the trials on which the same exact symptom pattern appeared (e.g. Gluck & Bower 1988). In function learning, for

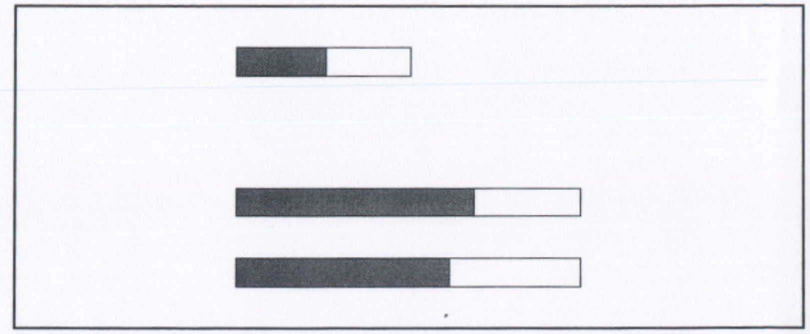


FIG. 11.1. Illustration of a display of stimulus, response, and criteria feedback for a typical function learning task. First the stimulus bar on the left is displayed, second the response bar in the middle is drawn by the subject, and third the criterion bar on the right is displayed for feedback.

example, infection severity may be proportional to body temperature plus sum random error.

The cue patterns used in category-learning tasks are usually constructed from a small set of binary-valued features (e.g. fever, present or absent; cough, present or absent). The cues used in function learning tasks are constructed from a small set of continuously valued dimensions (e.g. body temperature in centigrade; white blood cell frequency). However, dimensional stimuli have been used in categorization experiments (e.g. Homa 1984, Nosofsky 1986, Ashby & Gott 1988), so this is not the critical property for distinguishing between category- and function-learning tasks.

The responses used in category-learning tasks are usually limited to a small number of nominal categories (e.g. heart disease present or absent). The responses used in function-learning tasks are usually an equally spaced subset of a continuum of criterion magnitudes (e.g. percentage of arterial blockage). However, numbers could be used as category labels (e.g. disease severity levels 1, 2, and 3), and so this is not sufficient for distinguishing between category- and function-learning tasks.

The crucial property for distinguishing category- from function-learning tasks is the nature of the cue–criterion mapping. In category learning, a discontinuous mapping is made from stimuli to categories. In function learning, a continuous function is used to map cues to criteria (i.e. $z(t) = F[x(t)]$ where F is a continuous function). For example, a discontinuous map jumps up or down at some point in an abrupt manner, whereas a continuous map changes gradually at each point in a smooth manner.

Another important difference between category- and function-learning tasks is the way that performance is measured. In category learning, performance is based on the percentage of correct responses. But this would not work in function learning, because responses may be technically incorrect but highly accurate. For example, if the prediction is 78 per cent arterial blockage and the criterion is 79 per cent, then the response is technically incorrect but highly accurate. So in function learning, performance is based on the mean absolute error (MAE) between the subject's prediction and the criterion. (Another measure, called the achievement index, is the correlation between the subject's prediction and the criterion).

III. SUMMARY OF BASIC FINDINGS ON SINGLE-CUE FUNCTION LEARNING

Although function-learning tasks may involve multidimensional stimulus cues (e.g. predict a student's grade point average based on both verbal and math Scholastic Aptitude Test scores), the majority of theoretical work (Carrol 1963, Brehmer 1974, Koh & Meyer 1991) has been limited to single cue tasks (e.g. predict a student's grade point average based on the total SAT score). Accordingly, this chapter is limited to a review of single cue experiments (see Klayman 1988, Slovic & Lichtenstein 1971, for more comprehensive reviews).

The ten basic principles summarized below provide a partial ordering of the difficulty of learning various types of functions from experience. These ten principles are generalizations of well-established experimental results that have been replicated across a variety of conditions.

Principle 1: Continuous functional relations are learned faster than arbitrary categorical relations

See Carrol 1963, Sniezek & Naylor 1978. The mapping from a stimulus set to a criterion set can be formed in two different ways: One is to use a continuous function to form the stimulus–criterion pairs (e.g. using a quadratic function); the second is to use a jagged function (produced by erratic pairings) of the same stimuli and criteria. Thus the stimulus set is identical in both mappings and so is the response set. The only difference is the continuity of the mapping. In this comparison, continuous mappings are learned faster than erratic mappings of the same stimuli and criteria. So far, this result has been obtained with positive linear, negative linear, and quadratic functions, but the results may hold for other continuous functions.

Principle 2: Increasing functions are learned faster than decreasing functions

See Brehmer 1971, 1973, 1974, Brehmer et al. 1974, Naylor & Clark 1968, Naylor & Domine 1981. A function is increasing if its slope (derivative) is always positive, and it is decreasing if its slope is always negative. More specifically, researchers have compared performance for positive and negative linear functions, and they have found that positive linear functions are learned much faster. This finding, however, may not be restricted to linear functions.

Principle 3: Monotonic functions are learned faster than non-monotonic functions

See Carrol 1963, Brehmer 1974, Brehmer et al. 1985, Brehmer et al. 1974, Byun 1995, Deane et al. 1972, Delosh 1995, Sheets & Miller 1974, Sniezek & Naylor 1978. Monotonic functions always increase, or always decrease, but never do both. Non-monotonic functions increase and decrease at different cue values and are generally more difficult to learn. For example, Delosh (1995) has shown that both linearly decreasing functions and exponentially increasing functions are learned more quickly than non-monotonic quadratic functions.

Principle 4: Cyclic functions are more difficult to learn than non-cyclic functions

See Byun 1995. A cyclic function, such as a sine or cosine function, periodically changes directions producing a repeating increasing–decreasing pattern. Non-cyclic functions, such as a quadratic function, do not produce a repeating up–down pattern. To distinguish cyclic from non-cyclic functions within a finite range of cue values, cyclic functions are defined as functions that contain at least two repetitions of an up–down pattern (or two repetitions of a down–up pattern). At this point, only one experiment has been conducted comparing a sine function and a quadratic function, but the results demonstrated such a highly robust difference, that we believe it is safe to conclude that cyclic functions are much more difficult to learn that non-cyclic functions.

Principle 5: Linearly increasing functions are learned faster than nonlinearly increasing functions

See Byun 1995, Delosh et al. 1996. Increasing non-linear functions such as power, exponential, logarithmic, or logistic, always increase, but the rate of increase changes as a function of the cue value. Generally, these non-linear increasing functions are more difficult to learn than linearly-increasing functions. It is important to note that this conclusion depends on the psycho-physical scales used to measure the stimuli and criteria.

For example, Koh & Meyer (1991) found superiority for linear functions only after using a logarithmic scale to measure the continua. Thus the scales used to measure the continua is a key factor for determining the order of difficulty of learning linear versus power functions.

Principle 6: Predictions made at the beginning of training correlate with a linear function

See Sawyer 1991, Summers et al. 1969. When subjects are given a "neutral" cover story, and then they are trained to learn a non-linear function, their responses at the beginning of training correlate most highly with those expected from a linear function. As training progresses, the correlation with the non-linear function steadily grows and exceeds the linear function. This result has been observed with S-shaped logistic functions, but it is likely to hold for a wider class.

Principle 7: Congruent cue labels improve performance

See Byun 1995, Koele 1980, Miller 1971, Muchinsky & Dudycha 1974, 1975, Sniezek 1986, Adelman 1981. The description of the cues elicits prior knowledge about the functional relation that is either congruent, incongruent, or uninformative. For example, suppose subjects are asked to learn a relationship between x = price and y = quality of merchandise (suggesting a positive relation), but then they are trained with a negative linear function. In this case, the cue labels would be incongruent with the training function. Generally, performance is best with congruent labels, and worst with incongruent labels. However, even with incongruent labels, subjects gradually adjust and learn the appropriate cue–criterion relationship.

Principle 8: Systematic training sequences facilitate learning of difficult functions

Byun (1995) and Delosh (1995) trained subjects on a function using either a systematically-increasing sequence of stimulus magnitudes during training, or a randomly-organized sequence of the same magnitudes. Training sequence had no effect on positive linear functions, but it facilitated learning of non-monotonic quadratic functions and cyclic functions, with systematic sequences producing slightly superior performance.

Principle 9: Performance on interpolation test stimuli is almost as accurate as performance on training stimuli

See Carrol 1963, Koh & Meyer 1991, Delosh et al. 1996. During the transfer test phase no feedback is provided, and new cue values are presented that never appeared during training. New transfer cue values

that lie inside the range of training values are called *interpolation test stimuli*. On interpolation trials, subjects tend to choose new responses that fall in between the trained criterion values. Previous research with linear, power, exponential, and quadratic functions indicate that predictions on interpolation tests are almost as accurate as the training stimuli.

Principle 10: Subjects can extrapolate, but not as accurately as they interpolate

See Carrol 1963, Surber 1987, Delosh et al. 1996, Wagenaar & Sagaria 1975. An extrapolation test stimulus is a cue value that lies outside the range of the training values. Previous research with linear, exponential, and quadratic functions indicate that subjects generate extrapolation responses, that is responses outside the range of the training criteria values. Their extrapolations are in the appropriate direction with respect to the training function, however, these extrapolations do not come as close to the programmed function as interpolations.

Summary

The first five principles suggest the following tentative order for the difficulty of learning a functional relation from experience: cyclic > non-monotonic > monotonic decreasing > monotonic increasing > linear. Previous researchers have generally explained these findings in terms of prior knowledge or hypotheses about rules used to make predictions (Brehmer 1974, Sniezek 1986, Sawyer 1991). When standard instructions and cue labels are employed, subjects initially expect the cue–criterion relationship to follow a positive linear rule (Principle 6). However, these prior expectations can be modified by changes in prior instructions or by cue labels (Principle 7). The facilitation of learning by systematic as opposed to random stimulus sequences presumably results from the facilitation of hypothesis testing by using systematic sequences (Klayman 1988). Principle 10 has been used to argue that subjects learn abstract rules rather than simple stimulus–response associations (Brehmer 1974, Carrol 1963).

IV. COGNITIVE MODELS OF FUNCTION LEARNING

Theoretical requirements

The ten principles summarized earlier provide guidelines for constructing a model of function learning. However, additional general theoretical constraints must be met as well. First, the model must have the same learning power as humans. For example, a model that approximates all functions by a 3rd degree polynomial is insufficient, because it cannot approximate a cyclic function, which humans can learn (Byun 1995). Secondly, the model must have the same learning speed as humans. For example, a powerful non-linear hidden unit connectionistic network model that requires several thousand feedback trials to learn a simple linear relation is unreasonable because humans can learn this in much less than a hundred trials. Third, we wish to formulate a model of function learning that is consistent with category-learning theory. In other words, we seek a common theoretical explanation for category- and function-learning. Presumably, humans rely on a single common learning process to learn stimulus–response mappings, whether or not the mapping is continuous. There are two quite different approaches to the construction of a model of function learning: one is a *rule-based approach*, and the other is an *associative-learning approach*.

Rule-based Learning Approach

According to this approach (Brehmer 1974, Carrol 1963, Koh & Meyer 1991), the rules that subjects use to make predictions are represented by a linear combination of a basis set of functions:

$$y(t) = b_0 f_0[(t)] + b_1 f_1[x(t)] + b_2 f_2[x(t)] + \ldots + b_k f_k[x(t)] \quad \mathbf{(11.1)}$$

The most common choice for the basis set of functions is the polynomial basis, $f_k[x] = x^k$, but other bases are possible such as log polynomial, Fourier, Gaussian, or wavelet. The basis set must be sufficiently powerful to closely approximate all smooth continuous functions.

According to the rule-based approach, learning is represented by a search for the appropriate choice of coefficients ($b_0, b_1, \ldots, b_k$) to fit the training function $F[x]$. For example, Brehmer (1974) assumed a cubic polynomial basis, and he assumed that subjects test a linear hypothesis first, followed by a quadratic hypothesis, followed by a cubic hypothesis. Alternatively, Koh and Meyer (1991) assumed a log polynomial basis, and they assumed that subjects gradually adjust all of the coefficients ($b_0, b_1, \ldots, b_k$) in a trial by trial manner in the direction of minimizing a loss function.

One problem with these rule-based models is the lack of specification of the trial-by-trial search process. For example, Brehmer (1974) never specified exactly how hypotheses were rejected, nor how the parameters for testing a hypothesis were chosen. A similar shortcoming applies to the Koh & Meyer (1991) model. A second problem is that they do not extrapolate in the same manner as humans.

Delosh et al. (1996) examined the extrapolations produced by polynomial and log polynomial models for linear, quadratic, and exponential training functions, and found that these models failed to reproduce the same pattern of extrapolations as humans. For example, when trained with a negatively-accelerated increasing exponential function, these models generate non-monotonic relations at the upper end of the extrapolation region, contrary to the humans who continued to produce monotonic increasing relations in this region. A third problem is that they are not built from assumptions consistent with current research on category learning. These rule-based models were developed independent of research on category learning, and so they fail to explain category and function learning within a common theoretical framework.

In view of these limitations of rule-based models, the remainder of this chapter will focus on *associative-learning models* (ALMs). This is not to claim that rule-based models can be completely eliminated. We simply leave the question concerning the construction of a successful learning algorithm for them open for future research.

Associative-learning approach

The following associative-learning model is an extension of the artificial neural network model of Knapp & Anderson (1984) and the exemplar-based connectionistic model of Kruschke (1992). The latter model is currently a highly successful model of category learning (see Nosofsky & Kruschke 1992, for a rigorous evaluation). The main advantage of this model is that it is built from assumptions that are consistent with the major findings on category learning. Another advantage of this model is that it employs a simple yet powerful learning algorithm. ALM makes the following assumptions (see Figure 11.2).

Assumption 1. The physical stimulus, $x(t)$, produces a perceptual image represented by a *distribution of activation* across a set of n input nodes:

$$\{x_1, x_2, \ldots, x_i, \ldots, x_n\},\ x_1 < x_2 < \ldots < x_i < \ldots < x_n \quad \mathbf{(11.2)}$$

Each input node, x_i, corresponds to a real number representing a potential stimulus value, and the index i represents the rank order of the node value. When a cue value, $x(t)$, is presented, it activates input node x_i from the set of n input nodes according to a Gaussian similarity function:

$$a_i[x(t)] = 1 / exp\{ (x_i - \psi_x[x(t)]) / \sigma_x\}_2 \quad \mathbf{(11.3)}$$

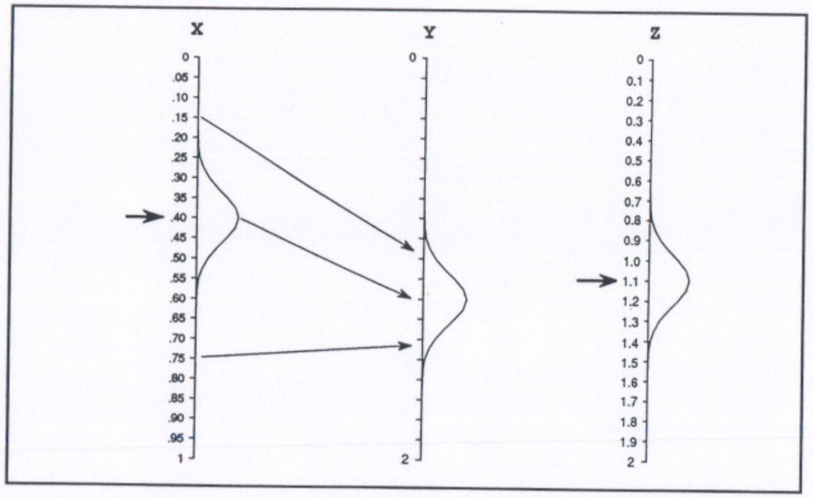


FIG. 11.2. Illustration of ALM. The perception of the stimulus is represented by the distribution of activation on the far left. This is mapped onto the output nodes by associations indicated by the arrows (only a few of the many are shown). The subjective response is represented by the distribution of activation shown in the middle. Finally, the perception of the criterion feedback is represented by the distribution of activation on the far right.

The psycho-physical function $\psi_x[x(t)]$ represents a subjective scaling of the physical stimulus (e.g. ψ is often approximated by a power function, $\psi(x) = \psi x^{\beta}$, see Stevens 1961). The parameter σ_x is used to determine the generalization gradient around each input node. For example, in Figure 11.2, the perception of the cue value $x(t) = 40$ is represented by a Gaussian distribution on the left centred at $\psi_x(40) = (0.01) \times 40 = 0.40$, with a standard deviation equal to $\sigma_x = 0.05$.

Assumption 2. The perceptual image of the criterion value, $z(t)$, is represented by a *distribution of activation* across a set of m criterion nodes,

$$\{z_1, z_2, \ldots, z_j, \ldots, z_m\},\ z_1 < z_2 < \ldots < z_j < \ldots < z_m \quad \textbf{(11.4)}$$

Each criterion node, z_j, is a real number, corresponding to a potential response category. When the criterion value, $z(t)$, is presented on trial t, it activates criterion node z_j according to a Gaussian similarity function:

$$c_j[z(t)] = 1 / \exp\{(z_j - \psi_z[z(t)]) / \sigma_z\}^2 \quad \textbf{(11.5)}$$

In Figure 11.2, the perception of the criterion value $z(t) = 110$ is represented by a Gaussian distribution on the right centred at $\psi_z(110) = (0.01) \times 110 = 1.10$ with a standard deviation equal to $\sigma_z = 0.05$.

Assumption 3. The subjective image of the response, $r(t)$, is represented by a *distribution of activation* across a set of m output nodes:

$$\{r_1, r_2, \ldots, r_j, \ldots, r_m\},\ r_1 < r_2 < \ldots < r_j < \ldots < r_m \quad \textbf{(11.6)}$$

The activation of output node, r_j, represents the subject's belief that category j is the correct response category. In Figure 11.2, the distribution of output activation in the middle is centred at a response node corresponding to a magnitude of 1.20.

Assumption 4. Each input node x_i is connected to each output node r_j by a weight $w_{ij}(t)$ representing the association between the pair of input and output nodes after t trials of training. The activation pattern distributed across the n input nodes is mapped by the $(m \cdot n)$ connection weights into an activation pattern distributed across the m response nodes. The response node r_j from the set of m response nodes is activated according to the linear associative map:

$$e_j[x(t)] = w_{1j}(t)a_1[x(t)] + w_{2j}(t)a_2[x(t)] + \ldots + w_{nj}(t)a_n[x(t)] \quad \textbf{(11.7)}$$

In Figure 11.2, the arrows indicate a few of the many associations from the inputs to the outputs.

Assumption 5. The distribution of activation across the m criterion nodes provides the feedback for updating the connection weights. The weight, $w_{ij}(t)$, connecting input node x_i to output node r_j is updated on trial t according to the following delta learning rule:

$$w_{ij}(t) = w_{ij}(t-1) + \alpha \cdot a_i[x(t)] \cdot \{c_j[z(t)] - e_j[x(t)]\} \quad \textbf{(11.8)}$$

Assumption 6. Prior knowledge is represented by the initial connection weights existing before training and evoked by task instructions, cue labels, and cover stories (cf. Choi et al. 1993). The initial connection weight between input node x_i and output node r_j, is symbolized $w_{ij}(0)$. Two different assumptions concerning the initial weights are examined in the simulations that follow. The first is called the *no-prior-knowledge-assumption*, which is obtained by setting $w_{ij}(0)$ equal to a value randomly sampled from a normal distribution with zero mean. The second is called the *positive-linear-prior-knowledge-*

assumption, which is obtained by teaching the network a positive linear relation prior to experience with the training function. For example, if the initial weights are set to $w_{ij}(0) = 1$ for $i = j$, and zero otherwise, then the network reproduces the identity function, $y = x$.

These six assumptions complete the description of the associative learning process. However, the assumptions concerning response selection have not been made explicit. At this point two different sets of assumptions are introduced: the first set describes a simple ratio rule for selecting responses, and the second set describes a more sophisticated linear interpolation extrapolation response rule.

Ratio response rule

According to previous category learning theories (e.g. Kruschke 1992), the response category is selected probabilistically according to a ratio rule. The probability of choosing output node r_j is assumed to be equal to:

$$Pr[r_j \,|\, x(t)] = e_j[x(t)] \,/\, \textstyle\sum_{k=1,m} e_k[x(t)]. \quad \textbf{(11.9)}$$

The output node r_j is a subjective scale value that needs to be located on the physical criterion scale to produce the observed response, $y(t)$, on trial t. Presumably the subject does this by choosing the physical response value that gives rise to the subjective image, r_j. Mathematically, this is equivalent to taking the inverse of the psycho-physical scaling function to produce the observed response:

$$y(t) = \psi_z^{-1}(r_j). \quad \textbf{(11.10)}$$

This implies that the mean prediction to cue value $x(t) = x_i$ equals:

$$\mu(x_i) = \textstyle\sum_{j=1,m} Pr[r_j \,|\, x_i] \cdot \psi_z^{-1}(r_j). \quad \textbf{(11.11)}$$

Linear interpolation–extrapolation response rule

Delosh et al. (1996) recently proposed a new model for response selection in function-learning tasks. The essential idea is that predictions are constructed from a linear interpolation–extrapolation rule (see, e.g. Figure 11.3).

First, the transfer test cue, $x(t)$, is matched to one of the previously experienced training values. For example, suppose that the new test cue, $x(t)$, is matched to the previous training value x_i. Consider the case where $x(t) < x_i$, and there is another training value x_{i-1} immediately below x_i. Then these two training values, $x_{i-1} < x_i$, are used to retrieve two outputs, $y(x_{i-1})$ and $y(x_i)$, respectively. Finally,

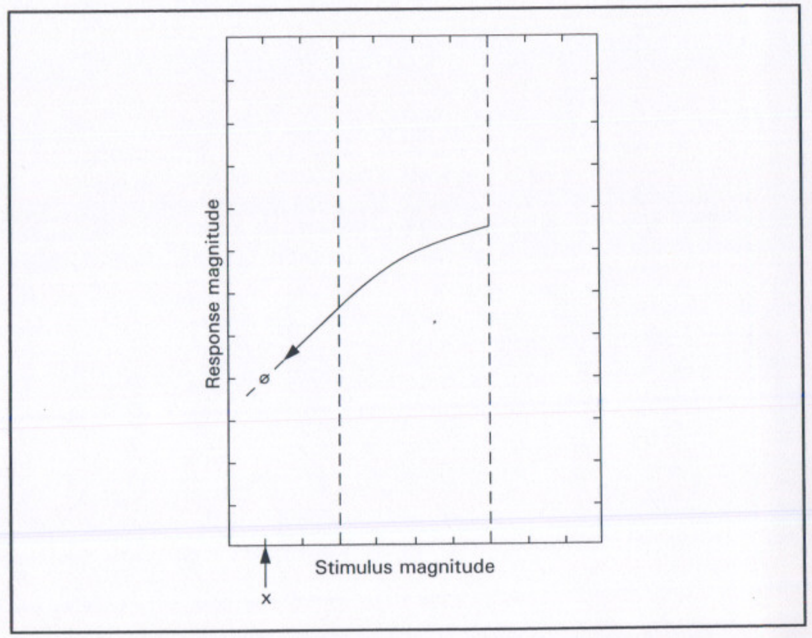


FIG. 11.3. Illustration of the linear interpolation–extrapolation rule. In this example, the training stimuli on the right are used to form a line that extrapolates down to the left to produce the prediction for the test stimulus indicated by the "x" on the horizontal axis.

the subject's prediction, $y(t)$, for the transfer test cue $x(t)$, is constructed by linear interpolation:

$$y(t) = y(x_{i-1}) + \{[y(x_i) - y(x_{i-1})] \,/\, (x_i - x_{i-1})\} \cdot [x_i - x(t)]. \quad \textbf{(11.12a)}$$

If $x(t) < x_i$, and there is no other training stimulus below x_i, then the prediction is linearly extrapolated:

$$y(t) = y(x_i) + \{[y(x_{i+1}) - (x_i)] \,/\, (x_{i+1} - x_i)\} \cdot [x(t)-x_i]. \quad \textbf{(11.12b)}$$

If $x(t) > x_i$, and there is another training stimulus x_{i+1} above x_i, then the prediction is linearly interpolated:

$$y(t) = y(x_i) + \{[y(x_{i+1}) - (x_i)] \,/\, (x_{i+1} - x_i)\} \cdot [x(t)-x_i]. \quad \textbf{(11.12c)}$$

If $x(t) > x_i$, and there is no other training stimulus above x_i, then the prediction is linearly extrapolated:

$y(t) = y(x_i) + \{[y(x_i) - y(x_{i-1})] / (x_i - x_{i-1})\} \cdot [x(t) - x_i]$ **(11.12d)**.

The probability of retrieving the output node $\psi[y(x_i)] = r_j$ using x_i as the retrieval cue is given by Equation 11.9, after substituting x_i for $x(t)$. The mean of $y(t)$ conditioned on matching $x(t)$ to x_i, denoted $\gamma(x_i)$, is obtained from Equation 11.12 by substituting $\mu(x_i)$ for $y(x_i)$ in the equation, where $\mu(x_i)$ is defined in Equation 11.11.

The new cue $x(t)$ is matched to an old training value x_i according to the following process. Training stimuli are identified by the learner as input nodes that produce a strong familiarity response. The familiarity of input node x_i is assumed to be determined by the maximum output activation produced by the input value $x(t)=x$:

$f_i(t) = max[e_1(x), e_2(x), ..., e_m(x)]$ **(11.13)**

(Note: Equation 11.13 is computed with $w_{ij}(0) = 0$ so that prior knowledge is not confused with training experience.) The strength of match of the new cue $x(t)$ to input node x_i, is determined by the product of familiarity and similarity:

$s_i[x(t)] = f_i(t) / exp\{(x_i - \psi_x[x(t)]) / \sigma_s\}^2$ **(11.14)**

where the similarity parameter, σ_s, is an unknown parameter. Finally, the probability of matching input cue $x(t)$ to input node x_i is determined by the ratio rule:

$p_i[x(t)] = s_i[x(t)] / \sum_{k=1,n} s_k[x(t)]$ **(11.15)**

This implies that the mean prediction to cue $x(t)$ is:

$E[y(t) \mid x(t)] = \sum_{i=1,n} p_i[x(t)] \cdot \gamma(x_i)$ **(11.16)**

In sum, the ratio-response rule uses Equations 11.9 and 11.11 to determine the mean prediction, while the linear interpolation-extrapolation response rule uses Equations 11.12, 11.13, 11.14, 11.15 and 11.16 to determine the mean prediction. The ALM in conjunction with the linear interpolation-extrapolation rule is called by Delosh et al. (1996) the *EXtrapolation Association Model (EXAM)*.

Simulation procedure

ALM involves only three unknown parameters: the generalization gradient for the physical continuum used to display the stimuli (σ_x in Equation 11.6); the generalization gradient for the physical continuum used to display the criteria (σ_z in Equation 11.7); and the learning rate (α in Equation 11.8). The same learning rate ($\alpha = 0.07$) and the same two generalization gradients ($\sigma_x = \sigma_z = 0.05$) were used in all of the simulations reported below. These parameter values were selected to reproduce all of the qualitative aspects of the basic findings, except for the results involving interpolation and extrapolation. The linear interpolation–extrapolation response rule requires an additional generalization gradient parameter (σ_s in Equation 11.14), which was set equal to $\sigma_s = 0.10$.

Several other specifications were necessary for the computer simulations. First, it was necessary to specify the psycho-physical functions, $\psi_x(x)$, $\psi_z(x)$. The simulations reported below were based on experiments that employed line lengths to physically display the stimuli and criteria. Past research has shown that the psycho-physical function for line lengths is approximately a linear function (at least within the limited range of magnitudes used in the experiments that were simulated). The following psycho-physical functions were used for all of the simulations:

$\psi_x(x) = x / [max(x) - min(x)]$
$\psi_z(z) = z / [max(x) - min(x)]$

where $max(x)$ represents the maximum cue value, and $min(x)$ represents the minumum cue value used in an experiment. This normalizes stimulus magnitudes so that $\psi_x(x)$ ranges from 0.0 to 1.0.

Next the input nodes were chosen to range from 0.0 to 1.0 in 0.01 step units {0.00, 0.01, 0.02, 0.03, . . ., 0.99, 1.0}, providing a dense coverage of the entire range of stimuli. To cover the criteria nodes, a wider range of nodes was employed: {0, 0.01, 0.02, 0.03, . . ., 2.99, 3.0}.

All of the simulations employed the same stimulus magnitudes and number of training trials as were used in the actual experiments. This was essential for reproducing the observed results. Most of the experiments employed randomly ordered stimulus sequences. Unless noted otherwise, the stimulus sequence used in the simulated training was also randomly ordered. The results were robust across different random orders.

V. REPRODUCING THE BASIC FINDINGS OF FUNCTION-LEARNING RESEARCH

The following presentation provides a constructive approach to model building. We begin with the simplest possible version of the ALM, and only introduce complexities as they are demanded by the data. This constructive approach is useful for identifying the importance of each new additional assumption of the model (see also Lamberts, this volume).

Principle 1

Carrol (1963) argued that associative learning models of function learning (such as ALM described earlier) can be ruled out because they fail to explain why continuous functional relations are easier to learn than arbitrary categorical relations. Carrol (1963) did not actually describe any specific associative-learning model, so this claim remains just a conjecture. It is elementary to prove that this conjecture is true for ALM when there is no generalization ($\sigma_x \to 0$). Thus it is interesting to see to what extent this criticism holds when generalization occurs across stimuli (e.g. $\sigma_x = 0.05$).

Carrol (1963) compared two groups of subjects: one trained with a continuous linear function, and another trained with an erratic function of the same stimuli and criteria as the linear function. Carrol (1963) also trained another two groups: one trained with a continuous non-monotonic quadratic function, and another trained with an erratic function of the same stimuli and criteria as the quadratic function.

ALM was trained on the same stimulus–criterion pairs and using the same number of training trials as used by Carrol (1963). The simplest possible version of ALM was employed. In particular, the ratio response rule was used to generate the model predictions and no prior knowledge was assumed. Table 11.1 shows the MAE, averaged across training, obtained from the human learners by Carrol (1963) in comparison with the computer simulation results obtained by ALM. The last two columns present a comparison of the continuous condition with the erratic control for linear and quadratic functions.

TABLE 11.1
Mean absolute error for each condition of Carrol (1963)

Function	*Continuous Condition*	*Erratic Control*
Human Data		
Linear	0.03	0.94
Quadratic	0.58	1.29
Simulated Data		
Linear	0.03	0.21
Quadratic	0.13	0.29

The human data was obtained by first computing the absolute error between each subject's prediction and the correction criterion value, and then averaging across subjects. The simulated data was obtained by first computing the average, and then computing the absolute error between the mean prediction and the criterion. The latter procedure produces smaller absolute errors by eliminating individual subject response variability.

As can be seen in Table 11.1, ALM reproduces the correct ordering of MAE across the four conditions. In particular, ALM reproduces the difference between the continuous and erratic functions. This results from the use of an adequate generalization gradient (e.g. $\sigma_x = 0.05$). If the generalization gradient is too narrow (e.g. $\sigma_x = 0.01$), then the ALM no longer produces any difference between the continuous and erratic conditions, just as Carrol (1963) claimed. Thus, the ALM with no prior knowledge, a ratio-response rule, and an adequate generalization gradient is sufficient for explaining Principle 1.

Principle 2

The first simulation did not employ any prior knowledge. However, this assumption fails to account for the fact that decreasing functions are more difficult to learn than increasing functions. If there is no prior knowledge, it is elementary to prove that ALM predicts no difference in rate of learning for decreasing as compared to increasing linear functions (when the same stimulus and criterion sets are used). Thus it is interesting to see to what extent that this problem can be eliminated by incorporating positive linear prior knowledge into the initial connection weights.

Naylor & Clark (1968) compared two groups of subjects: one trained with a positive linear regression equation: $z = 40 + 0.80 \cdot x + error$; and another trained with a negative linear regression equation: $z = 90 - 0.80 \cdot x + error$. The univariate distributions for the stimuli and the criteria were identical across the two conditions. The ALM was trained with the same stimuli and criteria using the ratio-response rule. The initial weights were set to reproduce the identity relation, $y = x$, before experiencing the training function.

Figure 11.4 shows the human results obtained from Naylor & Clark (1968) and Figure 11.5 shows the simulation results from ALM with prior knowledge. Each figure shows the achievement index (the correlation between the prediction and the criterion) plotted as a function of training block. The top curve in each figure represents the positive linear condition, and the bottom curve represents the negative linear condition. Thus, the ALM with adequate generalization, ratio-response rule, and positive linear prior knowledge is sufficient for explaining Principles 1 and 2.

Principles 3 and 4

Recall from the first simulation that even when there is no prior knowledge, the ALM predicts faster learning for linear as compared to quadratic functions (see Table 11.1). It is fairly obvious that the use of positive linear prior knowledge can only facilitate this advantage of

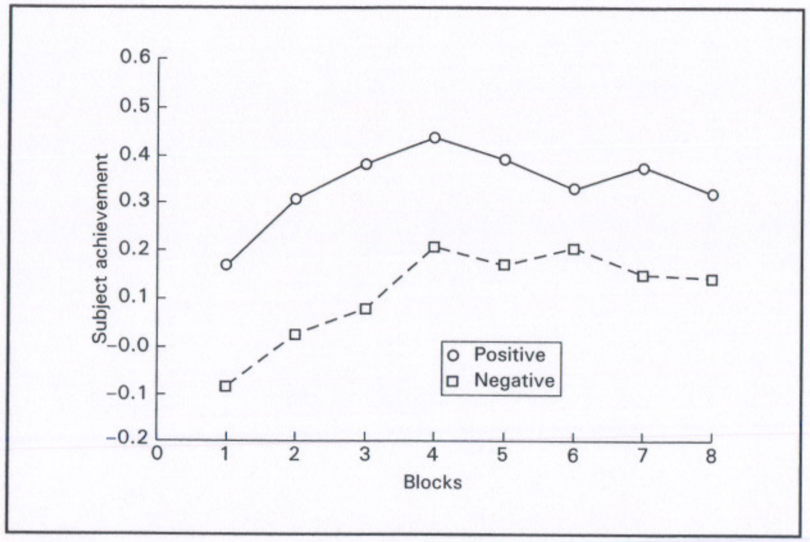


FIG. 11.4. Achievement index (correlation between prediction and criterion) plotted as a function of training for the positive and negative linear functions. Data from Naylor & Clark (1968).

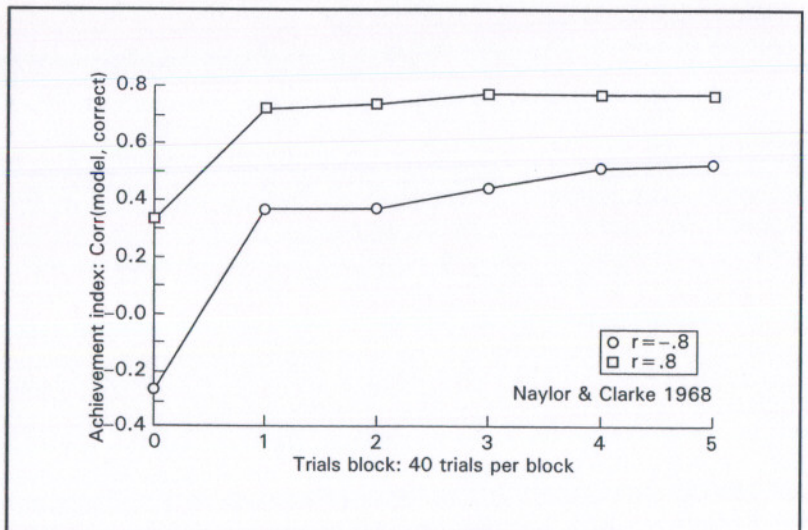


FIG. 11.5. Achievement index (correlation between prediction and criterion) plotted as a function of training for the positive and negative linear functions. Data from ALM simulation (Naylor & Clark 1968).

positive linear functions over quadratic functions. However, it is not clear that this prior knowledge will produce any advantage for quadratic functions over cyclic functions. Thus it is of interest to see to what extent the ALM can reproduce the differences between quadratic and cyclic functions.

Byun (1995) examined positive linear, non-monotonic quadratic, and cyclic functions as shown in Figure 11.6. The figure shows the criterion plotted as a function of the stimulus magnitude, with a separate curve for each training function. The ALM was trained on the same stimulus –criterion pairs for the same amount of training using the ratio-response rule. The initial weights were set to reproduce the identity relation, $y = x$, as in the previous simulation. The empirical results from Byun (1995) are shown in Figure 11.7, and the simulation results are shown in Figure 11.8. Each figure shows the MAE plotted as a function of training block. As can be seen by comparing Figures 11.7 and 11.8, the ALM reproduces the difficulty ordering: cyclic > non-monotonic > positive linear.

Principle 5

So far, the ALM has succeeded in reproducing the order of learning difficulty for functions that are categorically different in form (e.g. increasing versus decreasing, monotonic versus non-monotonic). A

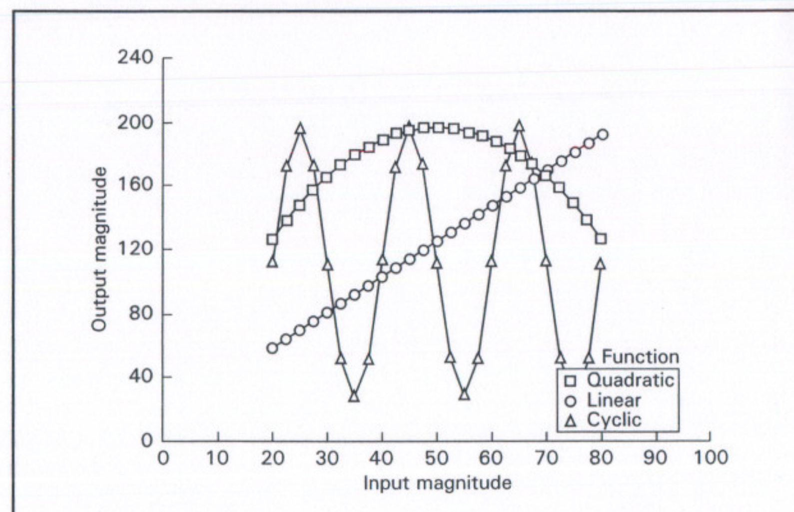


FIG. 11.6. Criterion plotted as a function of stimulus magnitude for the positive linear, quadratic, and cyclic functions examined by Byun (1995).

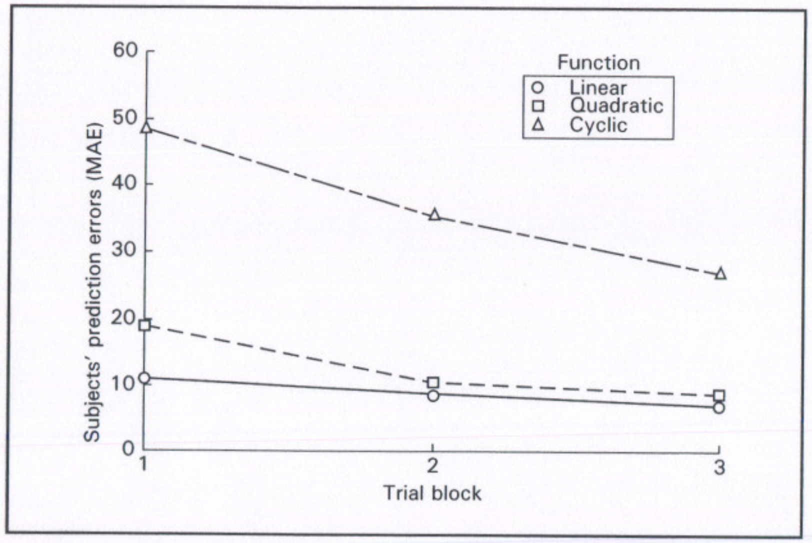


FIG. 11.7. Mean absolute error plotted as a function of training block separately for the positive linear, quadratic, and cyclic functions. Data from Byun (1995).

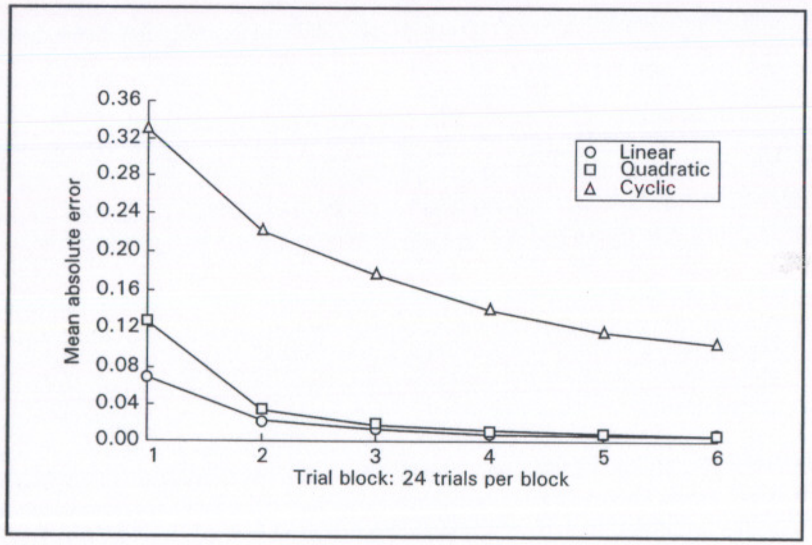


FIG. 11.8. Mean absolute error plotted as a function of training block separately for the positive linear, quadratic, and cyclic functions. Data from ALM simulation (Byun, Experiment 1).

greater challenge is to see whether or not ALM can reproduce the order of learning difficulty for functions that differ in more subtle quantitative forms. Byun (1995) compared five different monotonically increasing functions as shown in Figure 11.9: positive linear, negatively accelerated power, positively accelerated power, logarithmic, and logistic. The criterion is plotted as a function of the stimulus magnitude, with a separate curve for each of the five functions. The order of learning difficulty obtained from these five functions is shown in Figure 11.10a: logistic > logarithmic > positively accelerated power > negatively accelerated power > positive linear.

When the version of ALM described in the previous two simulations was applied to this data set, the model failed to reproduce the observed order. The main reason was the choice of prior knowledge. In the previous two simulations, the initial weights were set to reproduce the simple positive linear relation, $y = x$. This is a rather crude approximation, but it worked well enough for the categorically different functions that were examined in the previous simulations. However, to capture the subtle differences among the functions shown in Figure 11.9, it is necessary to select the initial weights more carefully. A better selection for the initial weights is to use a *proportional prior-knowledge* assumption, i.e. the minimum cue value is initially mapped onto the

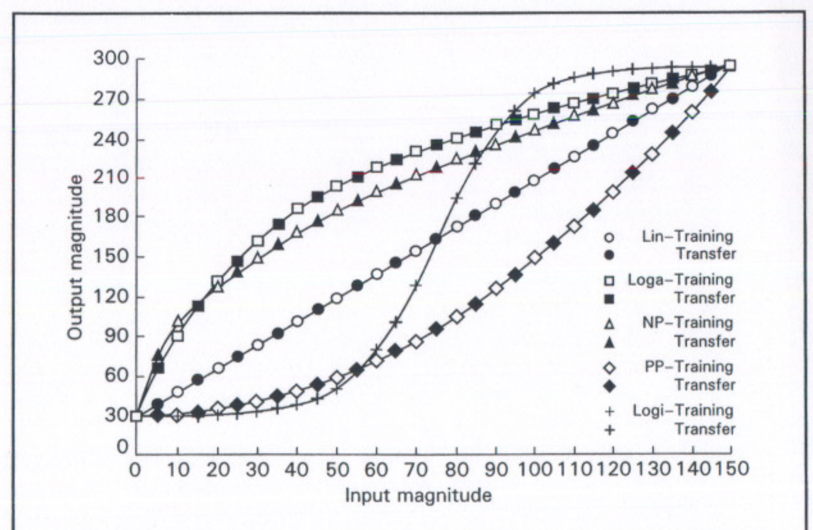


FIG. 11.9. Criterion plotted as a function of stimulus magnitude for the positive linear (PL), negatively accelerated power (NP), positively accelerated power (PP), logarithmic (LN), and logistic (LG) functions examined by Byun (1995).

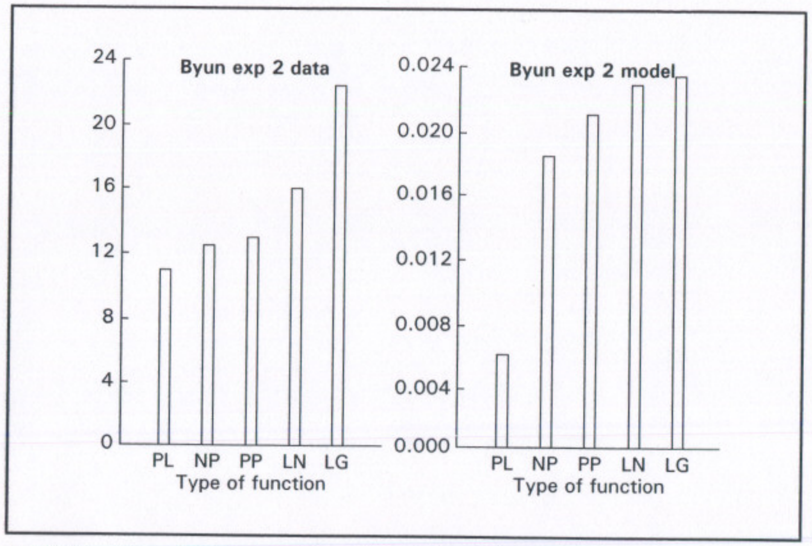


FIG. 11.10. Mean absolute error produced by each type of function. Panel a on left is data from Byun (1995). Panel *b* on right is data from ALM simulation. PL = positive linear, NP = negatively accelerated power, PP = positively accelerated power, LN = logarithmic, LG = logistic.

minimum criterion value, the maximum cue value is initially mapped onto the maximum criterion value, and intermediate stimuli are initially mapped proportionally as follows:

$$[y - min\,(z)]/[max\,(z) - min\,(z)] = [x - min(x)]/[max(x) - min(x)] \quad \textbf{(11.17)}$$

The ALM with this proportional prior knowledge assumption reproduces the observed order as shown in Figure 11.10b.

It is informative to examine the reasons why the logarithmic and logistic functions are so difficult to learn compared to the other functions. Figure 11.11 shows the mean predictions (averaged across subjects) as a function of stimulus magnitude produced by the human learners in comparison with the training function, as reported in Byun (1995). As can be seen in the figure, the logarithmic and logistic functions contain more curvature. The human learners underestimate concave (negatively accelerated) sections of these functions, and overestimate convex (positively accelerated) sections of the function. Figures 11.12a and 11.12b shows the predictions produced by ALM in comparison with the training function, for the logarithmic and logistic functions. As can be seen in Figures 11.12a and b, ALM produces the same pattern of errors

– overestimating concave and underestimating convex. This is because of the fact that the generalization gradient of the ALM causes the model to produce an interpolated prediction.

Principles 6 and 7

As illustrated in the three previous simulations, prior knowledge is directly built into the model by the selection of the initial weights. Many so called "neutral" cover stories and cue labels tend to evoke initial weights that conform to the positive linear prior-knowledge assumption. However, cover stories and cue labels can be constructed that evoke quite different initial expectations or initial weights. For example, if subjects are asked to learn a relation between x = sedative amount and z = patient activity, then this would tend to evoke a negative linear set of initial weights, such as for example, $w_{ij}(0) = 1$ if i = -j, and zero otherwise. In

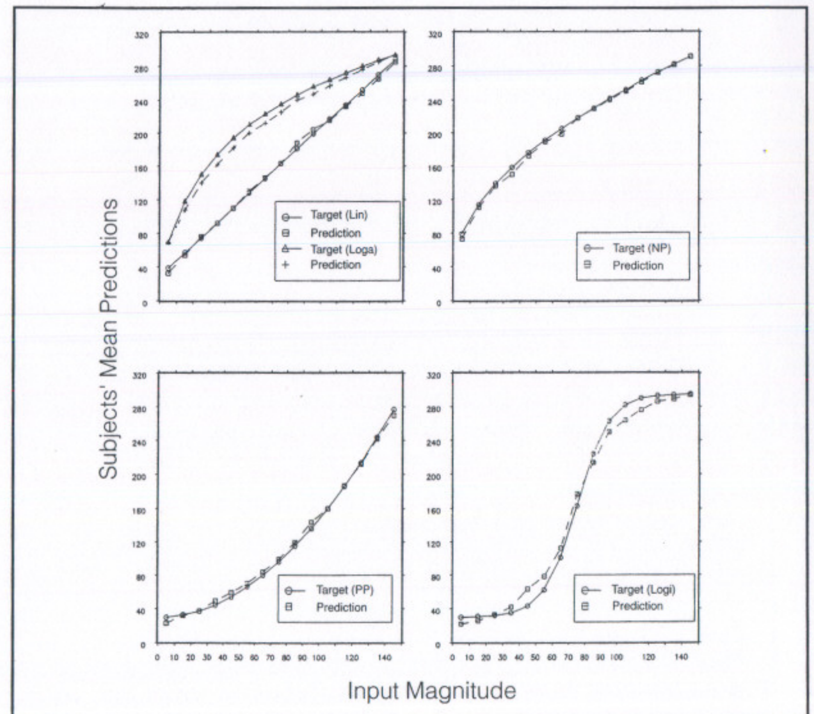


FIG. 11.11. Mean of subjects' predictions plotted as a function of stimulus magnitude for the logarithmic (top left), negatively accelerated power (top right), positively accelerated power (bottom left), and for the logistic function (bottom right). The training function criterion values are also plotted in each figure.

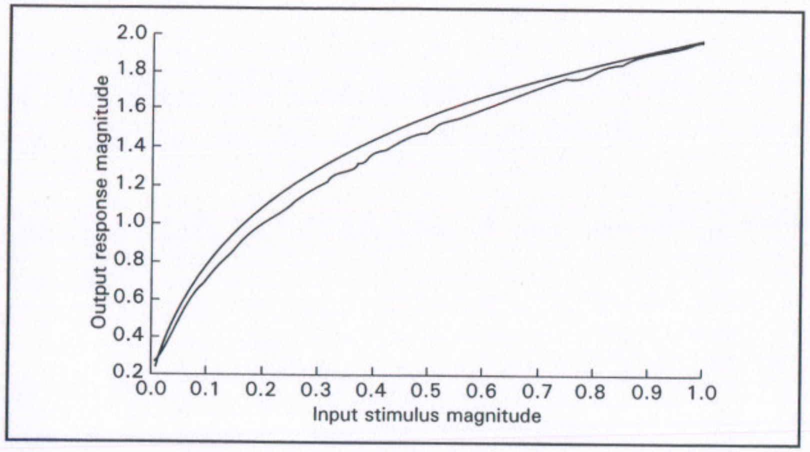


FIG. 11.12a Predictions of ALM plotted as a function of stimulus magnitude for the logarithmic functions.

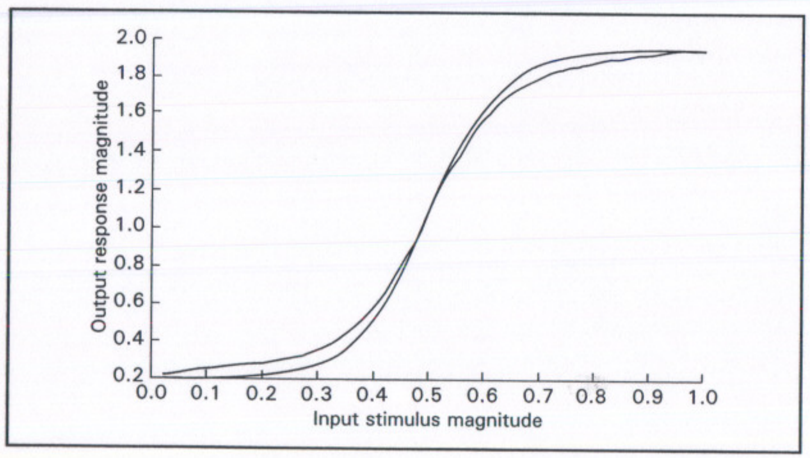


FIG. 11.12b Predictions of ALM plotted as a function of stimulus magnitude for the logistic function.

this case, the learning curves shown in Figure 11.5 would be produced, except that the top curve would now represent the negative linear training condition (congruent with the prior knowledge).

Principle 8

The effect of systematic as compared to random training sequences provides a very strong challenge to the ALM because there is no specific

mechanism in the model designed to produce such effects. Thus it is quite interesting to see whether or not the ALM can account for the improvement in learning produced by systematic sequences.

Delosh (1995) investigated the order of learning difficulty produced by negative linear as compared to non-monotonic quadratic functions. In addition, he examined the effects of training with systematic versus random stimulus training sequences. The basic results were that negative linear functions were easier to learn than quadratic functions, and furthermore systematic sequences produced better performance than random sequences. The ALM was trained on the same stimulus magnitudes and training sequences as used by Delosh (1995), using the response-ratio rule and the positive linear prior-knowledge assumption. The MAE for each type of function and training sequence produced by the simulation are shown in Table 11.2. As can be seen in the table, the ALM yields better performance with negative linear as compared to quadratic functions, and also there is an advantage produced by training ALM with systematic as compared to random sequences.

The systematic training advantage for ALM is a generally important demonstration. One might not expect that artificial neural networks would be influenced by the organization of the training sequence. Indeed, the advantage of systematic over random training sequences has heretofore been assumed to implicate a hypothesis testing process of function learning. It is now clear that training sequence effects can emerge as well from associative learning processes.

Principles 9 and 10

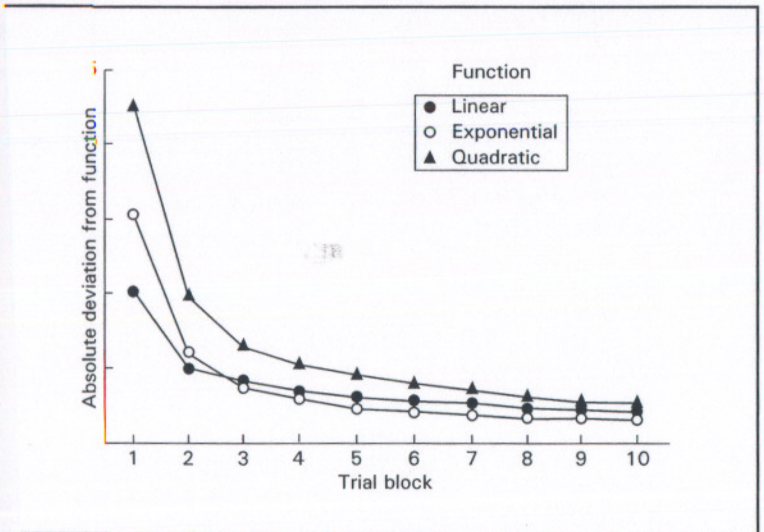
A number of theorists (Carrol 1963, Brehmer 1974) have argued that the strongest evidence favouring rule-based models over associative-learning models is obtained by examining extrapolation performance. Abstract rules provide systematic guidelines for extrapolating beyond experience, whereas simple associations between stimuli and criteria experienced during training provide no mechanism for extrapolation outside the range of experience (Delosh et al. 1996). This criticism may not apply to ALM because it allows for generalization on both the stimulus and criteria continua, thus it is of interest to see the extent to which ALM can account for interpolation and extrapolation performance.

TABLE 11.2
Mean absolute error produced by ALM for each condition of Delosh (1995)

Function Form	*Random Sequence*	*Systematic Sequence*
Negative Linear	0.033	0.028
Quadratic	0.044	0.031

Deloet al. (1996) trained subjects on the middle range of stimulus magnites for linear, exponential, and non-monotonic quadratic functio: Following this training, they later tested subjects on new interpoion test stimuli (new values inside the training range), and new extpolation test stimuli (new values outside the training range). The ALI'as trained on the same stimulus magnitudes and training trials ased by Delosh et al. (1996), using the ratio response rule, and using iral weights that reproduced the simple identity relation ($y = x$). Figu11.13 shows MAE plotted as a function of training produced by the han subjects, and Figure 11.14 shows the corresponding plot producey the ALM. Once again, the ALM reproduced the observed order of learn: difficulty (quadratic > exponential > positive linear).

Figur11.15 and 11.16 illustrate the predictions of the ALM for the positiveear training condition. Note that at this point, the predictions are basen the ratio-response rule. Figure 11.15 shows what happens when theneralization gradient is too tight (e.g. $\sigma = 0.01$), and Figure 11.16 iltrates the results for a wider generalization gradient e.g. $\sigma = 5$). The top straight line in both figures represents the linear rainingnction ($z = .3 + 2.2 \cdot x$, using the normalized stimulus scale). The bott straight line in both figures represents the prior knowledge dentity ation ($y = x$). The jagged line in Figure 11.15 represents the



i. 11.13. h absolute error plotted as a function of training block separately for the positive ar, exponei and quadratic functions. Data from Delosh et al. (1996).

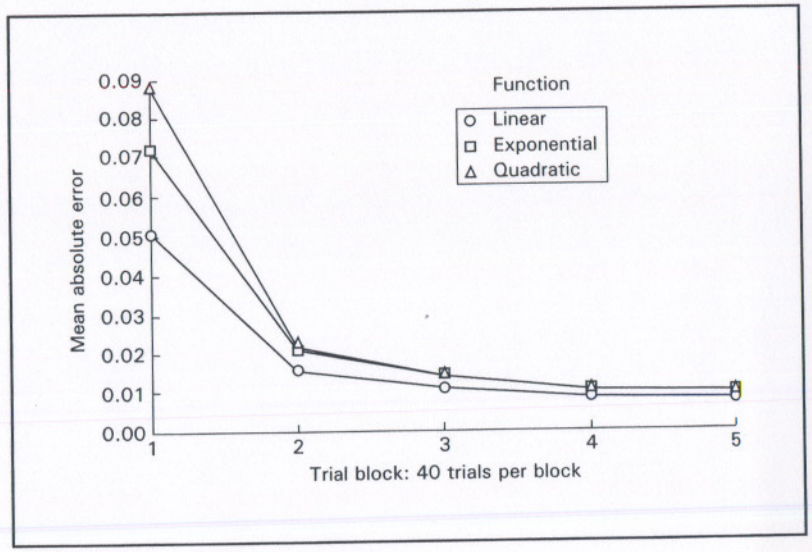


FIG. 11.14. Mean absolute error plotted as a function of training block separately for the positive linear, exponential, and quadratic functions. Data from ALM simulation (Delosh, Experiment 1).

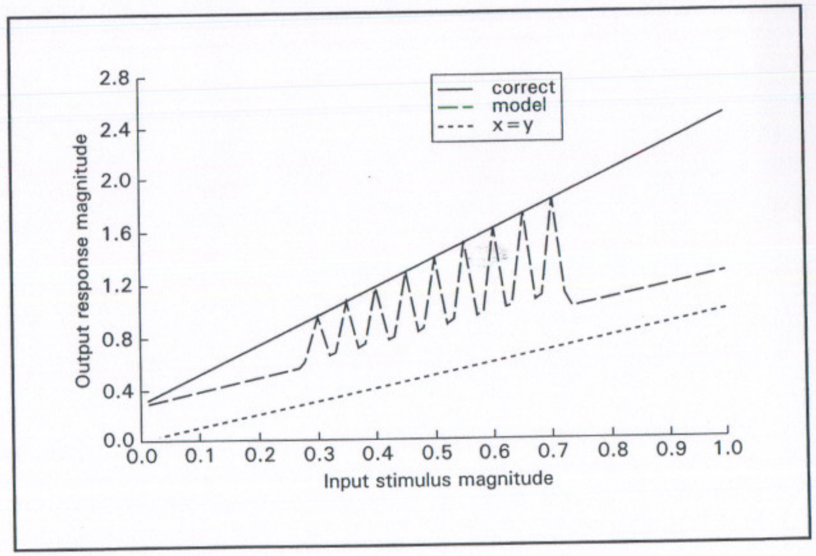


FIG. 11.15. Predictions produced by ALM plotted as a function of stimulus magnitude for the linear function using a tight generalization gradient.

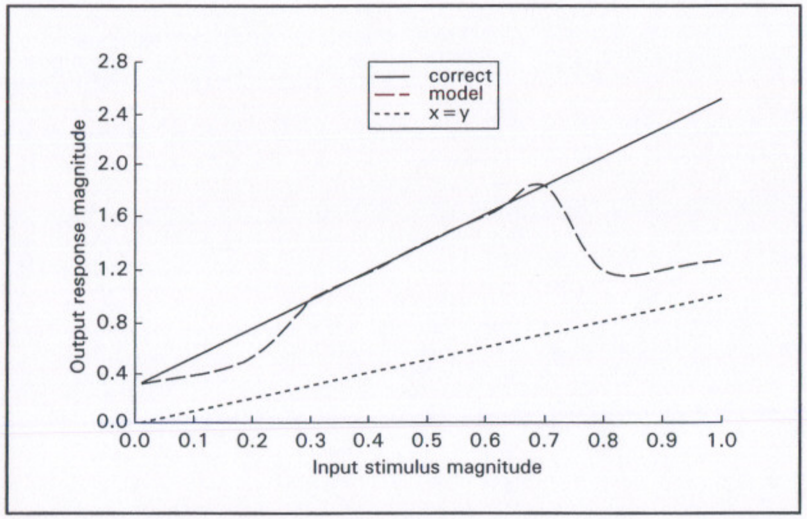


FIG. 11.16. Predictions produced by ALM plotted as a function of stimulus magnitude for the linear function using a wider generalization gradient.

predictions generated by ALM with no generalization. The training stimulus values produce the predictions by ALM that peak and intersect with the training function criterion values. The sudden drops above and below each peak indicate a failure of the ALM to interpolate when there is no generalization. The smooth curve in Figure 11.16 represents the predictions of ALM with generalization. Now ALM produces appropriate interpolation responses, but it still fails to extrapolate appropriately in the extreme lower and upper transfer test regions – here the predictions of ALM fall back toward the prior knowledge identity function.

In summary, if a ratio-response rule is employed along with a sufficiently wide generalization gradient, then ALM can interpolate but it cannot extrapolate. Thus the criticism of associative learning models by rule-based theorists appears to be partly right. But not entirely, because ALM can be salvaged by adding the linear interpolation–extrapolation response rule. Figure 11.17 illustrates the extrapolation performance by humans in the left panel, and by EXAM in the right panel for the exponential function condition. Figure 11.18 shows the corresponding results produced by a polynomial rule-based model (Equation 11.1). As can be seen in these figures, the extended version of ALM (called EXAM) provides a better account of the pattern of human extrapolation than the rule-based models (see Delosh et al. 1996, for more details).

CONCLUSIONS

The purpose of this chapter was to begin building a bridge between theoretical work on category learning and function learning. Category learning and function learning appear to be closely related, and it seems useful to determine the extent to which it is possible to formulate a common theoretical explanation for both domains of research. Towards this aim, an artificial neural network model originally developed for category learning was extended to make it applicable to function learning. The most important extension was the addition of a linear interpolation–extrapolation response rule. With this extension, the model represents a hybrid or integrated approach to associative and rule-based models. Prior knowledge and learning are represented by simple associations, but rules are evoked during test to construct a sophisticated response from simple associations.

Several important ideas were discovered from this theoretical endeavour. First, artificial neural network models need to carefully

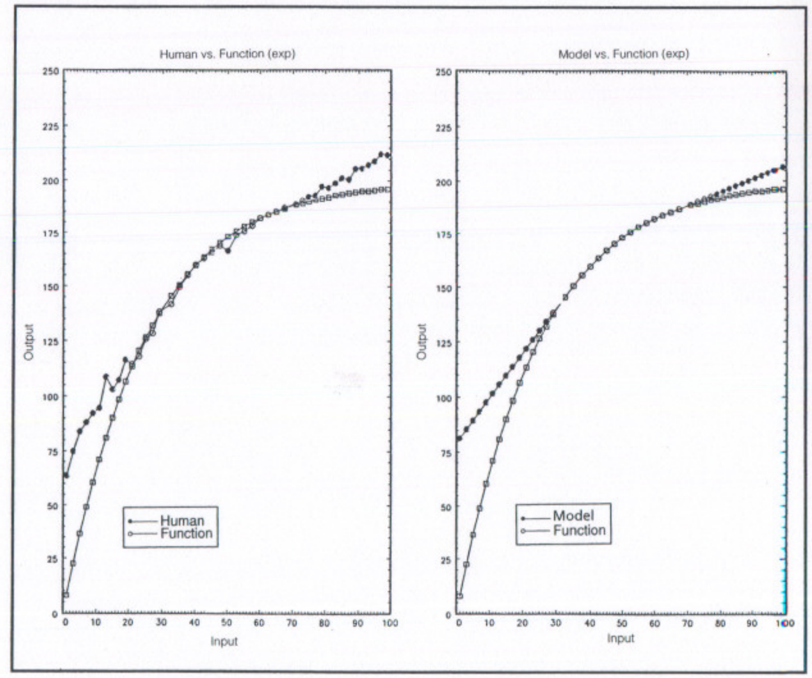


FIG. 11.17. Predictions from humans (left) and EXAM (right) plotted as a function of stimulus magnitude for the exponential training function.

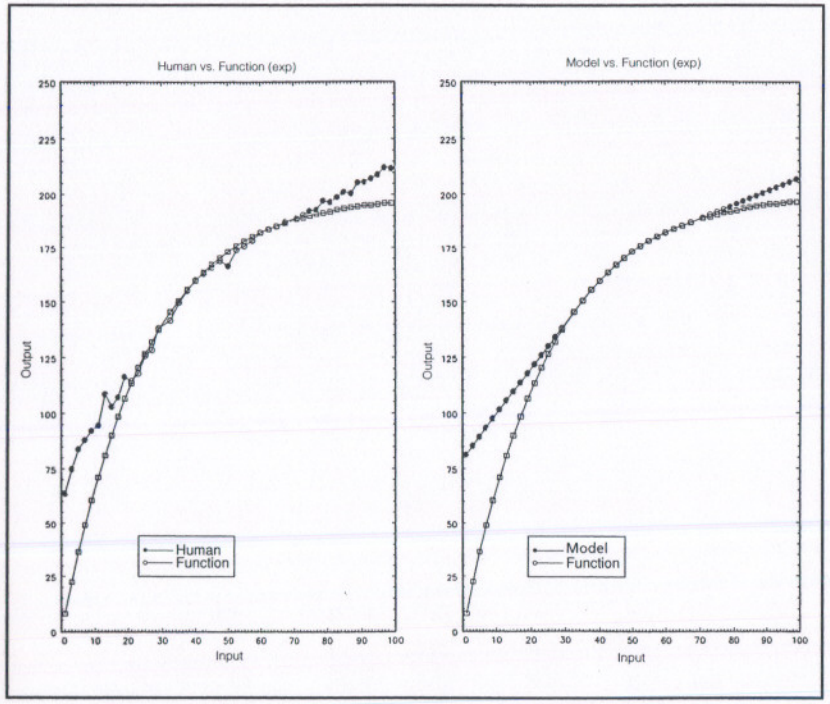


FIG. 11.18. Predictions from humans (left) and polynomial rule model (right) plotted as a function of stimulus magnitude for the exponential training function.

match the initial associations of the network with the prior knowledge evoked by the task instructions. For example, prior knowledge was necessary to account for the fine grain differences in the difficulty of learning various types of non-linear increasing functions. Secondly, artificial neural network models need to use sophisticated rules to construct responses from the output information retrieved by the network. For example, a linear response rule based on retrieved outputs was necessary to account for extrapolation behaviour in function learning. Thirdly, artificial neural network learning models are sensitive to the organization of the training sequence. For example, a model trained with systematically-increasing stimulus magnitudes produced faster learning than a model trained with randomly-ordered stimulus magnitudes.

Ten basic principles of function learning were synthesized from the empirical literature, and we evaluated the extent to which the model could reproduce these principles. We were successful in finding a

common set of model parameters that reproduced all ten principles. (This does not imply that the predictions are insensitive to parameters; on the contrary, the predictions vary dramatically as a function of the generalization gradients.) We conclude that the model provides an excellent starting point for generating simple and parsimonious reproductions of the basic facts from both category and function learning.

REFERENCES

Adelman, L. 1981. The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance* **27**, 423–42.

Ashby, F. G. & R. E. Gott 1988. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**, 33–53.

Brehmer, B. 1971. Subjects' ability to use functional rules. *Psychonomic Science* **24**, 259–60.

Brehmer, B. 1973. Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior and Human Performance* **9**, 377–95.

Brehmer, B. 1974. Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance* **11**, 1–27.

Brehmer, B., H. Alm, L. E. Warg 1985. Learning and hypothesis testing in probabilistic inference tasks. *Scandinavian Journal of Psychology* **26**, 305–13.

Brehmer, B., J. Kuylenstierna, J-E. Liljergren 1974. Effects of function form and cue validity on the subjects' hypothesis in probabilistic inference tasks. *Organizational Behavior and Human Performance* **11**, 338–54.

Byun, E. 1995. *Interaction between prior knowledge and type of nonlinear relationship on function learning*. PhD dissertation, Psychology Dept., Purdue University.

Carrol, J. D. 1963. *Functional learning: the learning of continuous functional mappings relating stimulus and response continua*. Research Bulletin RB–63–26, Educational Testing Service, Princeton, New Jersey.

Choi, S., M. A. McDaniel, J. R. Busemeyer 1993. Evaluation of adaptive network models of conceptual rule learning. *Memory and Cognition* **21**, 413–23.

Deane, D. H., K. R. Hammond, D. A. Summers 1972. Acquisition and application of knowledge in complex inference tasks. *Journal of Experimental Psychology* **92**, 20–26.

Delosh E. L. 1995. *Hypothesis testing in the learning of functional concepts*. Masters thesis, Psychology Dept., Purdue University.

Delosh, E. L., J. R. Busemeyer, M. A. McDaniel 1997. Extrapolation: The *sine qua non* of abstraction. To appear in *Journal of Experimental Psychology: Learning, Memory, Cognition*.

Estes, W. K. 1994. *Classification and cognition*. New York: Oxford University Press.

Gluck, M. A. & G. H. Bower 1988. From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General* **117**, 227–47.

Homa D. 1984. On the nature of categories. In *The psychology of learning and motivation* (vol. 18), G. H. Bower (ed.), 49–94. New York: Academic Press.

Klayman, J. 1988. On the how and why (not) of learning from outcomes. In *Human judgement: the SJT approach*, B. Brehmer & C. R. B. Joyce (eds), 115–62. Amsterdam: North Holland.

Knapp, A. & J. A. Anderson 1984. A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning, Memory, Cognition* **10**, 616–37.

Koele, P. 1980. The influence of labelled stimuli on nonlinear multiple-cue probability learning. *Organizational Behavior and Human Performance* **26**, 22–31.

Koh, K. & D. E. Meyer 1991. Function learning: induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, Cognition* **17**, 811–36.

Kruschke, J. K. 1992. alcove: an exemplar-based connectionistic model of category learning. *Psychological Review* **99**, 22–44.

Miller, P. M. 1971. Do labels mislead? A multiple cue study, within the framework of Brunswick's probabilistic functionalism. *Organizational Behavior and Human Performance* **6**, 480–500.

Muchinsky, P. M. & A. L. Dudycha 1974. The influence of a suppresser variable and labelled stimuli on MCPL. *Organizational Behavior and Human Performance* **12**, 429–44.

Muchinsky, P. M. & A. L. Dudycha 1975. Human inference behaviour in abstract and meaningful environments. *Organizational Behavior and Human Performance* **13**, 377–91.

Naylor J. C. & R. D. Clark 1968. Intuitive inference strategies in interval learning tasks as a function of magnitude and sign. *Organizational Behavior and Human Performance* **3**, 378–99.

Naylor, J. C. & R. K. Domine 1981. Inferences based on uncertain data: some experiments on the role of slope magnitude, instructions, and stimulus distribution shape on the learning of contingency relations. *Organizational Behavior and Human Performance* **27**, 1–31.

Nosofsky, R. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**, 39–57.

Nosofsky, R. & J. K. Kruschke 1992. Investigations of an exemplar-based connectionistic model of category learning. In *The psychology of learning and motivation* (vol. 28), D. L. Medin (ed.), 207–50. New York: Academic Press.

Sawyer, J. E. 1991. Effects of risk and ambiguity on judgments of contingency relations and behavioral resource allocation decisions. *Organizational Behavior and Human Performance* **49**, 124–50.

Sheets, C. A. & M. J. Miller 1974. The effect of cue-criterion function form on multiple cue probability learning. *American Journal of Psychology* **87**, 629–41.

Slovic, P. & S. Lichtenstein 1971. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance* **6**, 649–744.

Sniezek, J. A. 1986. The role of variable labels in cue probability learning tasks. *Organizational Behaviour and Human Decision Processes* **38**, 141–61.

Sniezek, J. A. & J. C. Naylor 1978. Cue measurement scale and functional hypothesis testing in cue probability learning. *Organizational Behaviour and Human Decision Processes* **22**, 366–74.

Summers, S. A., R. A. Summers, V. T. Karkau 1969. Judgments based on different functional relations between interacting cues and criterion. *American Journal of Psychology* **82**, 203–11.

Surber, C. F. 1987. Formal representations of qualitative and quantitative reversible operations. In *Formal operations in developmental psychology: progress in cognitive development research*, J. Bisanz, C. J. Brainerd, R. Kail (eds). New York: Springer.

Wagenaar, W. A. & S. D. Sagaria 1975. Misperception of exponential growth. *Perception and Psychophysics* **18**, 416–22.