Psychological Research (2000) 64: 105–116 © Springer-Verlag 2000

ORIGINAL ARTICLE

Michael L. Kalish · John K. Kruschke

The role of attention shifts in the categorization of continuous dimensioned stimuli

Received: 20 November 1998 / Accepted: 25 August 1999

Abstract Results of human category learning experiments, using stimulus dimensions with binary values, have implicated a rapidly acting mechanism of attention shifts. Theories of categorization desire that stimuli with binary, discrete and continuous valued dimensions should all be treated similarly. Theoretical analyses of attention shifting, however, have up to now only been developed for shifts between features, or shifts between entire dimensions, not shifts within dimensions. Here we present a model of how people learn to discriminate categories made up of stimuli with continuous-valued dimensions. The model uses rapid shifts in attention within stimulus dimensions to reduce errors during learning; the model generalizes J. K. Kruschke's (*Psychological Review, 99,* 22–44, 1992) ADIT model. In an experiment in category learning, subjects were trained to discriminate four bivariate normal distributions that are presented with differential base rates. The base-rate manipulation produces several qualitative effects, for which the model accounts very well. With attention shifting turned off, the model fails to account for some aspects of the data, suggesting that attentions shifts are an important mechanism in the model.

Introduction

An object that is to be categorized (''stimulus'') is composed of a number of dimensions, and variation among stimuli can be either qualitative or quantitative. Qualitative differences include the difference between a patient with a runny nose, versus one with a cough, or the difference between an animal that flies and one that swims. Such features have been termed ''substitutive'' (Tversky, 1977), because any single object can only have one of the two possible nominal values. Qualitative differences also include cases where the object to be categorized either has some nominal feature, or it doesn't, such as a patient who might or might not have a runny nose. Such features are called ''present/absent'' (Gluck & Bower, 1988). Finally, stimulus variation can be genuinely quantitative, where the value can be measured, and ordered according to magnitude. Quantitative dimensions, can, of course, be either continuous or discrete, and ratio, interval or ordinal scaled. Many stimulus dimensions used in psychological research are continuous, with color (Nosofsky & Palmeri, 1996), length (Kalish & Kruschke, 1997), and angle (Ashby & Maddox, 1992) being common examples.

Models of category learning have been developed for all three of these stimulus types, but are often challenged when set to predict the results of differential category base rates. The relative frequency of contrasting categories during training typically alters what people learn about the stimuli within each category. Base-rate effects have been found for different types of stimuli: substitutive, present/absent (Gluck & Bower, 1988; Medin & Edelson, 1988), and continuously-valued (Healy & Kubovy, 1981; Maddox, 1995).

Attention has frequently been claimed to play a role in category learning; however, attentional theories have been developed only for shifts between present/absent (binary-valued) dimensions (e.g., ADIT, Kruschke, 1996) or for shifts between continuous dimensions (e.g., ALCOVE, Kruschke, 1992), but not for shifts within dimensions.

We conceive stimulus dimensions as being psychologically represented by a large set of elements, with a particular value of the dimensions being represented by the activation of a subset of those elements. This is analogous to the classic stimulus sampling theory of Estes (1950), except that we also think of the elements as being ordered when the dimension is psychologically ordinal (and our formal model does not incorporate

M. L. Kalish
Department of Psychology, University of Western Australia, Nedlands, WA 6009, Australia

J. K. Kruschke
Indiana University, Bloomington, Indiana, USA

random sampling of elements in its present instantiation). The important principle we add to this classic psychological conception is the notion that each element has its own attention strength, which can be shifted in response to error. For any given stimulus value, the learner has the option of differentially attending to its various elements. For example, consider two similar values, which activate overlapping sets of elements. If these values must be discriminated because they belong to different categories, then it would be beneficial if the learner shifted attention away from the shared elements toward the distinctive elements. The formal model we introduce later implements this simple psychological principle of attention shifts among elements.

We report new empirical research regarding the effects of base rates on category learning in continuous dimensions, and we report new theoretical work regarding an implementation of rapid attention shifting, *a la* ADIT, for continuous dimensions. The article first describes the inverse base-rate effect, and an explanation of the effect in terms of the ADIT model. Then we describe how the principles in ADIT can be used in a model, called CORNER, that handles continuous-valued dimensions. We test the model with data from an experiment that attempts to detect the influence of attention shifts within continuous-valued dimensions. We show that CORNER fits the data well when its attentional shifting mechanism is intact, but that it deviates systematically from the data when its attention shifting mechanism is excluded.

An example of rapid attention shifts: the inverse base-rate effect

When people learn to categorize stimuli drawn from categories that have unequal base rates, their responses differ from what normative statistical theory would predict. For a given stimulus, it is possible to compute the posterior probability of each of the categories. If subjects were to respond at the optimal rate, they would always choose the category label with the highest posterior probability; choosing the most likely category always maximizes the long-run rate of correct responses. People (and other animals) typically do not maximize in this way but, instead, match their response frequencies to the posterior probabilities, a behavior called "probability matching". Sensitivity to base rates can be found when subjects' response frequencies accord with the posteriors as computed by Bayes' theorem. Conversely, when the posteriors and the responses diverge, the base rates have had a non-normative effect on behavior. More simply, if subjects are shown an uninformative stimulus after training, their responses ought to match the actual base rates, but this is usually not the case.

In certain cases, subjects are likely to choose the less likely (rare) category more often as the more likely (frequent) one, even though the posteriors are equal. The equality of the posteriors depends on the normative use of base-rate information, the effect is termed "apparent base-rate neglect" (Gluck & Bower, 1988). Similar effects are seen in the categorization of a continuous stimulus dimension: when two categories are presented with unequal base rates, subjects tend to respond with the common choice less often than the posteriors predict, over a range of stimulus values. This effect has been termed "base-rate conservatism" (Green & Swets, 1967; Healy & Kubovy, 1981). Perhaps the most extreme base-rate effect has been found with present/absent features: in the "inverse base rate" of Medin and Edelson (1988), the difference between the posteriors actually favored the more frequently presented category, but people actually choose the rare category label more often than the frequent category label.

Let us consider the inverse base-rate effect more closely. Here, stimuli have three dimensions with present/absent features.

Actually, *pairs* of common and rare categories are defined over three dimensions. Since three such pairs were used in the original experiment, there were nine dimensions, in total, in the study. For stimuli in class 'F', the dimensions take on the values (1, 1, 0). For stimuli in class 'R', the values are (1, 0, 1). Class F stimuli occur three times as often as class R, making F frequent and R rare. The second dimension is a perfect predictor of the frequent class (call it 'PF'). The third dimension (PR) is a perfect predictor of the rare class, and the first dimension (I) is an imperfect predictor, occurring with both response classes. The inverse base-rate effect occurs when subjects are shown the stimulus (0, 1, 1) that contains both the PF and the PR dimensions: they respond with the rare category label more often than with the frequent category label, even though the base rates make the frequent category statistically more likely. Paradoxically, when subjects are shown the stimulus (1, 0, 0), they are more likely to respond with the frequent category.

Explaining the inverse base-rate effect in terms of attention shifting

Recently, both the inverse base-rate effect and apparent base-rate neglect have been explained by a single mechanism (Kruschke, 1996). Interestingly, the explanation involves the base rates only tangentially and, instead, uses the concept of rapidly shifting dimensional attention. According to this model (ADIT), base-rate effects are caused by associative learning and shifting attention. Early in training, people learn to associate the dimensions I and PF with the response class F, because F stimuli are presented more often. Subsequently, when an R stimulus is presented, the I dimension signals class F, because that association has already been learned. The I dimension thus produces an error because the stimulus is actually from class R. Subjects reduce the error by directing their attention away from dimension I, leaving their attention mostly on dimension PR. Sub-

jects, therefore, associate dimension PR with class R. Because two dimensions are associated with F and one dimension is associated with R, the weight connecting PR to R is greater than the weight connecting PF to F. The asymmetry leads to the inverse base-rate effect when the model is presented with the stimulus (0, 1, 1) after training, and classifies it as an example of class R. Kruschke (1996) predicted that the attention-shifting explanation would cause a pseudo inverse base-rate effect under *equal* base rates, when, for example, category F was learned before presentations of R began. This prediction was confirmed experimentally.

Old and new models of rapid attention shifts: ADIT and CORNER

Attention plays a central role not only in Kruschke's theory of category learning (Kruschke, 1996), but also in Nosofsky's generalized context model (GCM) of categorization (Nosofsky, 1986). The GCM has been extremely successful, accounting for categorization of discrete and continuous stimuli, drawn from arbitrary distributions, and presented in a large number of different physical forms (see Nosofsky, 1998). The main competitor to the GCM is general recognition theory (GRT) (Ashby & Gott, 1988; Ashby & Maddox, 1992; Ashby & Townsend, 1986). While the GCM and GRT are closely related formally (Ashby & Alfonso-Reese, 1995; Marley, 1992), there are substantial conceptual differences; for example, GCM assumes that only exemplar information is used during categorization, whereas GRT holds that only abstract rules are used during categorization. Most relevant here, GRT assumes that attention plays only a minor role in categorization and category learning. In the GRT, attention to a stimulus dimension reduces the level of perceptual noise on the attended dimension. However, GRT provides no mechanism to allow attention to change during a single trial or between trials. Thus, the GRT assumes that learning and attention are separate systems. GCM also does not include a mechanism to link learning and attention, but the model has been extended to include one; Kruschke's (1992) ALCOVE theory added an attention learning mechanism to the GCM. In ALCOVE, errors made during learning cause attention to stimulus dimensions to change slowly over time. Neither the GRT, the GCM or ALCOVE, however, can account for the base-rate effects on category learning that were described above.

ADIT's guiding principle is that attention and learning are tightly connected but are not identical. In ADIT, categorization errors during learning are used to re-allocate attention among stimulus attributes within a single trial. The attention shifts yield long-term associations that cause contrasting categories to be represented differently from each other, e.g., the first-learned category will be associated with all its relevant stimulus attributes, whereas the second category will be associated only with its distinctive features. Although the verbal description of the theory underlying ADIT makes general reference to stimulus attributes, ADIT itself operates only in a domain of stimuli with present/absent dimensions. Thus, the mechanisms of ADIT represent each stimulus feature as a separate stimulus dimension. Here, we present a model, based on ADIT, that can handle continuous-valued dimensions (we call the model CORNER, an acronym that reflects its generalization of ADIT) and test the model against data from a categorization experiment. CORNER includes ADIT as a special case, when each dimension has only two values. We first present ADIT formally, and then develop the CORNER model.

Formal description of ADIT

ADIT is a two-layer network, consisting of a layer of input units and a layer of output units. The model takes as input a stimulus x that is made up of d binary-valued dimensions, $\{x_1 ... x_i ... x_d\}$. Each input unit represents a single dimension, while each output unit represents a single category. The activation of the input unit representing dimension i is given by:

$$a_i^{in} = \begin{cases} 1 & \text{if } x_i \text{ is present} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Each input node is connected to every output node. Each input node has an attention strength, and each input-output connection has an associative weight. These two factors jointly determine the influence of each input feature on each of the k category node activations:

$$a_k^{out} = \sum_i \alpha_i w_{ik} a_i^{in} \quad (2)$$

where α_i is the attention given to feature i, and w_{ik} is the weight from feature i to category node k.

Following the presentation and encoding of the stimulus, every active input unit draws an equal measure of attention to itself. The initial allocation of attention depends only on the number of features present in the stimulus; adjustments made to attention on previous trials do not carry over into later trials. Because attention is a limited resource, its allocation is normalized over all active units:

$$\alpha_i := \frac{\alpha_i}{\left(\sum_i \alpha_i^{\eta}\right)^{1/\eta}} \quad (3)$$

where η is a free parameter ($\eta > 0$) that represents the level of attention available to the network.

When the input nodes are activated, and attention has been distributed across them, the activation of each output unit can then be computed, using Equation 2. The network then receives corrective feedback, identifying the correct response category. Error is computed by comparing the activation of each output unit with a "humble" teacher signal (Kruschke, 1992). Humble

teachers allow the output unit to over-shoot the target value without correction:

$$t_k = \begin{cases} \max(1, a_k^{out}) & \text{if x is in category } k \\ \min(0, a_k^{out}) & \text{otherwise} \end{cases} \quad (4)$$

The error for each of the k category nodes is set to $E_k = t_k - a_k^{out}$, and the total error is

$$E = 0.5 \sum_k {E_k}^2 \quad (5)$$

Although both attention strengths and association weights are adjusted to reduce the total error, these adjustments occur sequentially. Attention is first shifted to those features of the stimulus that produce the least error. Only after attention has been re-allocated (and re-normalized) does the model re-compute the category node activations and adjust the association weights on the basis of the new activation levels.

Attention strengths are shifted away from input nodes that contribute more error toward nodes that contribute less error, by moving attention in the direction opposite to the gradient of error:

$$\begin{aligned} \Delta\alpha_i &= -\lambda_\alpha \frac{\delta E}{\delta \alpha_i} \\ &= \lambda_\alpha a_i^{in} \sum_k E_k w_{ik} \end{aligned} \quad (6)$$

where λ_α, is a free parameter, termed the attention-shift rate.

After attention has been adjusted, negative values are set to zero. Attention is then re-normalized using Equation 3. Output activations and error are re-computed, so that the long-term association weights can then be adjust to further reduce error:

$$\begin{aligned} \Delta w_{ik} &= -\lambda_w \frac{\delta E}{\delta w_{ik}} \\ &= \lambda_w \alpha_i a_i^{in} E_k \end{aligned} \quad (7)$$

where λ_w is a scaling constant, termed the learning rate.

To evaluate the predictions of the model, overt response probabilities are computed in two steps. First, the information from the stimulus is considered, and then the base-rate information is factored in. Stimulus information comes from mapping the category node activation onto unbiased response probabilities. The probability that the model will choose category q is

$$p_q = \frac{\exp(\phi a_q^{out})}{\sum_k \exp(\phi a_k^{out})} \quad (8)$$

where k is the number of category nodes. This exponentiated form of the Luce choice rule (Luce, 1963) is chosen for several reasons. First, it allows the model to reflect the level of certainty subjects have in their response selection. When ϕ is large, small differences in category node activations are mapped to large differences in response probabilities. When ϕ is small, it takes large differences in activation to produce any change in response probabilities. Second, because category node activations are allowed to be negative, exponentiation maps activations onto non-negative response probabilities. Finally, exponentiated mapping has been used in previous models of categorization (Estes, 1994; Gluck & Bower, 1988; Kruschke, 1992).

Like the similarity-choice rule (Nosofsky, 1988), the choice probabilities in ADIT are biased; unlike the similarity-choice rule, the bias is not a free parameter of each response category. In ADIT, response biases depend only on the actual category base rates, which are weighted relative to other stimulus information. The base-rate estimates for each of the k categories, b_k, are accurate estimates of the true category relative frequencies during training. The response bias is applied with a weight proportional to a constant, β that reflects the value that subjects place on base-rate information relative to stimulus-specific information:

$$p_q^i = \frac{p_q b_q^{\beta/\beta+N}}{\sum_k p_k b_k^{\beta/\beta+N}}, \quad (9)$$

where N is the number of stimulus dimensions presented. When N is large, base-rate information is used proportionally less than when N is small, reflecting the greater amount of information present in a stimulus with more dimensions.

Formal description of CORNER

To extend ADIT, changes were made in both the representations used in the model, and in the method for distributing initial attention.

In ADIT, each input dimension is represented as a single value, whereas in CORNER each input dimension is represented psychologically as a vector of values. As with ADIT, CORNER is a linear network – the output values are (as shown below) simply the weighted sum of the input values.

The stimuli used in the experiment below varied in height. The input representations were chosen to reflect the fact that the dimensions used in the experiment were marked, with 'tall' the marked end of the dimension. As discussed in the introduction, the inherent asymmetry in a marked dimension means that a symmetric representational scheme, such as that used by Kalish and Kruschke (1997) will not work. Rather than introducing a mechanism for stimulus sampling, we instead use "thermometer" encoding (Anderson, Silverstein, Ritz, & Jones, 1977), which we describe next.

The input to the CORNER network is the set of psychological values for a multidimensional stimulus $x = \{x_1 ... x_i ... x_d\}$. Each dimension is represented by a set of hidden nodes, so that the activity of any given node j along dimension i is determined by x_i, the value of the stimulus on dimension i. The thermometer activation function depends on the location of the input node

within the vector that represents dimension i, and the magnitude of the stimulus on dimension i:

$$a_{ij}^{in} = \begin{cases} 1 & \text{if } (x_i - \mu_{ij}) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where x_i is the magnitude of the stimulus on dimension i and μ_{ij} is the location of node j of dimension i. The stimulus magnitude is greater than the location of the input node, then the node is fully active. When the stimulus value is less than the location of the input node, the node is completely inactive. The radical asymmetry means that tall and short stimuli are represented quite differently in the network; tall stimuli are represented by a large number of active input units relative to short stimuli.

The activation rule in ADIT is a special case of thermometer coding. If the vector of input units for dimension i contains only one unit and the location of the unit is such that the unit will be active whenever the stimulus has a measurable value on dimension i, the unit will also be *inactive* whenever the value on dimension i is zero, as in Equation 1. Thermometer encoding effectively discretizes the continuous stimulus. Each discrete interval forms a 'feature', over which attention and associative learning can operate.

Following the encoding of the stimulus, each input node then draws attention to itself:

$$\alpha_{ij} = a_{ij}^{in} \quad (11)$$

where α_{ij} is the attention devoted to unit j of dimension i. Like ADIT, attention is distributed across all active nodes, but in CORNER there may be more than one node for each dimension.

Once attention has been allocated, it is then normalized to reflect overall capacity limits. For each unit j of the vector of input units representing dimension i:

$$\alpha_{ij} := \frac{\alpha_{ij}}{\left(\sum_i \sum_k \alpha_{ik}^{\eta}\right)^{1/\eta}} \quad (12)$$

As with ADIT's Equation 3, η determines the amount of attention available. When η is very large, all units receive complete attention (i.e., $\alpha_{ij} \approx 1$), and when η is very small all units receive very little attention. The value of η is estimated during model fitting.

Once attention has been normalized, the category predictions of the model can be made, just as in ADIT. The activation of each category output node is the weighted sum of the product of the activation and attention vectors. In CORNER, the activation of any output node k is given by:

$$a_k^{out} = \sum_i \sum_j \alpha_{ij} w_{ijk} a_{ij}^{in} \quad (13)$$

where the w_{ijk} are the association weights connecting input node j of dimension i to output node k. This is essentially the same activation rule as ADIT's Equation 2, but summed over the extra index of the number of nodes used to represent each dimension. In ADIT there is only one node per dimension.

Following activation of the category output nodes, the correct response is presented to the network, in the form of a humble teacher value (Equation 4).

Error is then computed by taking $E_k = t_k - a_k^{out}$ for each category node. Using the method of gradient descent on error, the attention strengths are updated to reduce the prediction error on the current stimulus. The change in attention is given by:

$$\Delta\alpha_{ij} = \lambda_\alpha a_{ij}^{in} \sum_k E_k w_{ijk} \quad (14)$$

which is identical to Equation 6 except for a change in subscripts.

Shifting attention due to error is the central concept of CORNER; Fig. 1 shows a small example of the shift graphically. The shifting of attention results in increased activity of some feature units, and the decreased activity of others, as in ADIT. Notice that CORNER shifts attention within dimensions, whereas ADIT only shifted attention between dimensions.

Once attention has been shifted to reduce error, output activations are recomputed as in Equation 13. The associative weights are then adjusted to reduce any remaining error,

$$\Delta w_{ijk} = \lambda_w a_{ij}^{in} \alpha_{ij} E_k \quad (15)$$

where λ_w is a free parameter.

To evaluate the fit of the model, overt response probabilities were computed in two additional steps, directly analogous to those taken in ADIT. First, the basic probability of choosing the category represented by a given node, q, is computed by Equation 8. Finally, the base-rate information is used to bias the response probabilities, in the manner given by Equation 9.

Experiment: Learning categories with continuous-valued dimensions

ADIT was originally formulated to describe the psychological processes underlying human performance in the learning of categories with different base-rates. Although the central mechanism of rapid attention shifts makes no reference to the base rates of the categories, ADIT describes the data because it captures the fact that people learn the more frequently presented categories before the infrequently presented categories. Attention shifts ensure that the representation of the second category is based on distinctive, rather than characteristic, features. The model uses base-rate information only to bias the response probabilities to all stimuli, regardless of what features they have. While ADIT successfully accounted for subjects' performance with discrete-valued dimensions, the question remains as to whether the generalized form of the attention-shifting construct used in CORNER can model the behavior of subjects in categorizing stimuli with continuous-valued dimensions.

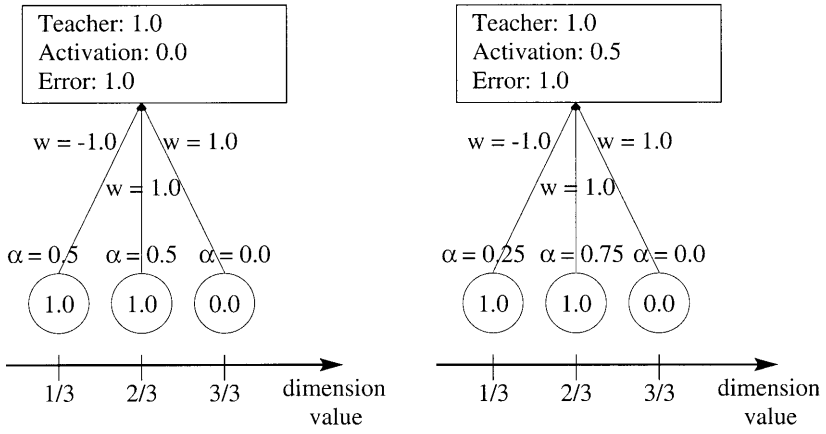

Fig. 1 The rapid shift of attention, as implemented in CORNER. The three input nodes represent consecutive values on a dimension. The stimulus has a value of 2/3 on this dimension. *Left panel* An initial presentation of a one-dimensional stimulus activates the input layer of processing nodes, which then attract attention. The attention is normalized to a fixed capacity, and category units are activated by multiplying the activation and attention of each input unit by the weight connecting it to each category unit. *Right panel* Error at the output level is backpropagated to adjust attention so as to reduce error. Because of differences in the long-term association weights between inputs and outputs, some input units are responsible for more error than others. Attention is diverted away from these units, and toward units that are responsible for less error. The new attention values are used to re-compute the output activations, and from them, response probabilities. Finally, the remaining error is used to adjust the long-term association weights (not shown)

Design

The goal of this experiment is to create "exceptional" dimension values for the rare categories. Figure 2 shows the design schematically. Four categories are defined as bivariate normal distributions, with equal variances. Two of the categories (the 'frequent' categories) are presented three times as often as the other two, 'rare' categories. In one experimental condition, the frequent categories have either tall or short values on both dimensions, in the alternate condition the two frequent categories each have one tall and one short value. In the design, each frequent category shares one mean value with one rare category, and its other mean value with the other rare category. CORNER predicts that learning will lead to asymmetrical associations and, thus, to response probabilities that are shifted with respect to the actual posterior probabilities. Any bias due to base rates alone cannot produce such shifts when the categories to be discriminated have the same central tendencies.

If subjects use base-rate information to match their response probabilities to the posterior probabilities of the categories in this experiment, they will choose the frequent categories three times more often than the rare categories at all points along both scales. Inverse base-rate effects would be seen if the rare categories are chosen more than the frequent ones, base-rate neglect or conservatism will appear if the rare category is chosen more than 1/4 of the time, and base-rate 'overuse' would be evidenced by less than 1/4 of responses being the rare category. Thus, base rates in and of themselves can produce only overall changes in relative response frequency, not local changes in relative response rates.

However, if subjects allocate their attention based on categorization errors, as CORNER predicts the results should be more subtle. Frequent stimulus values that are learned first should cause incorrect responses for infrequent stimuli learned later, and attention should become more focused on extreme stimulus values. Simulation modeling of CORNER, compared with a version of the model in which the rate of attention shifting is set to zero, follows the presentation of the results below.

To evaluate the predictions of CORNER, a large number of independent responses must be collected from each subject. Thus, following training, subjects are given a transfer test, in which stimuli are presented without corrective feedback. Because response frequencies must be measured at a large number of values along

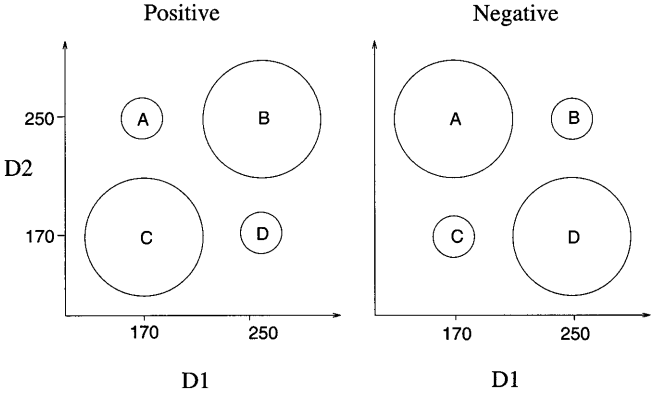

Fig. 2 Equiprobability contours for the stimulus distributions (categories) for the experiment. Two frequent and two rare categories were composed by drawing from bivariate normal distributions at four locations. The distributions have means located at the centers of the circles. All distributions had equal variance, and zero covariance (thus, they are circles), but they differed in their base rates, and so the circles had different radii. Physical stimuli instantiated this abstract structure with bars of one color having lengths given by dimension one, and bars of another color having lengths given by dimension two

each stimulus dimension, presenting bivariate stimuli during transfer would be very laborious. Instead, we presented subjects with single-dimensional stimuli in the transfer phase.

Method

Participants. Sixty undergraduates from Indiana University volunteered for partial course credits.

Apparatus. Subjects were tested on PC-type computers in individual, sound-dampened rooms.

Procedure. Each subject read instructions explaining the task, asking questions of the experimenter as needed. On each of 600 training trials, a stimulus consisting of two vertical bars (one red, one blue) appeared on the computer screen, along with a prompt to choose one of four possible responses. Each bar presentation had a random vertical offset, so that the height of the bar and the location of the top of the bar were not perfectly correlated. After the subject entered their response via the keyboard, the correct category label was presented, and the entire display remained visible for 1 s before the next trial began. A single set of stimuli was drawn from normal distributions and was presented to each subject in a different random order. The four distributions, A–D, were located at the corners of a square in the input space: A = (170, 250), B = (250, 250), C = (250, 170), D = (170, 170) (all units are pixels). All distributions had a nominal standard deviation of 35, but sampling resulted in centroids and deviations slightly different from these defining values. For one group of 30 subjects, categories A and C were frequent (the "Negative" group, because A and C lie on the diagonal with negative slope), presented three times as often as categories B and D. For the other 30 subjects, in the "Positive" group, categories B and D, which lie on the positive slope, were the frequent categories. Assignment of logical categories to response keys, and stimulus dimensions to physical bars, was randomly varied across subjects.

Testing. Following the training block, 58 test trials were presented to each subject. The test trials presented each stimulus dimension (bar) separately, taken uniformly from the range (13, 377). The most extreme transfer trials were thus outside the range of training trials, which ranged from 74 to 363 pixels.

Results

Training

The proportion of correct responses to each category in the last 100 trials of training was computed, and the two frequent categories and two rare categories were averaged into a frequent and rare proportion correct, respectively. As shown in Fig. 3, the Negative group learned the two frequent categories better than the Positive group, $F(1, 58) = 85.8$, $MSE = 0.0124$, $p < 1.0^{-6}$. However, the Positive group learned the two rare categories better than the Negative group; $F(1, 58) = 4.21$, $MSE = 0.0253$, $p = 0.045$. Thus, both the symmetric categories (those where stimuli were either short on both dimensions, or tall on both dimensions, labeled B and C in Fig. 2) were more difficult to learn than the asymmetric

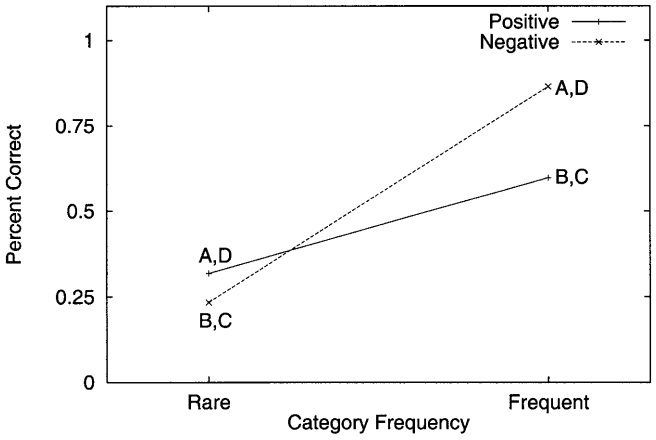

Fig. 3 The results of the training phase, with the percent correct for rare and frequent categories, for the Positive and Negative training conditions. Point labels refer to category names as shown in Fig. 2

categories, where stimuli (labeled A and D in Fig. 2) were tall on one dimension, and short on the other.

The responses during the last 100 trials of training were aggregated over subjects within each group to form observed response probabilities. Because the category structures were symmetric across the two axes, responses could be averaged over the two dimensions, to produce average responses for each stimulus value.

The category labels, however, cannot form the basis for response averaging, because individual categories occupy different locations on the two dimensions (such as category A, which is generally short on dimension 1 and tall on dimension 2). Therefore, four logical categories were constructed out of the four labeled categories. The four categories are defined by the position of their mean on a dimension (tall or short) and by their frequency: the frequent short (FS) category in the Positive condition is composed of "C" responses on both dimension 1 and dimension 2, and the rare short (RS) category in the Positive condition is composed of "A" responses on dimension 1 and "D" responses on dimension 2. The frequent tall (FT) consists of the "B" responses on both dimensions and the rare tall category (RT) consists of the "D" responses on dimension 1 and the "A" responses on dimension 2. For the Negative condition, the same construction is used to generate "tall" and "short" categories, but the relative frequencies were reversed, as indicated in Fig. 2.

Each stimulus presented during training contained both stimulus dimensions. The computation of response frequencies counts each trial twice, with each dimension separately contributing a logical category response. The likelihood measure of homogeneity is reduced by half to reflect the dependence of shared scores on each trial. Responses were collected into bins at 50 pixel intervals. Figure 4 shows the probabilities of each of the logical category responses for the Positive and Negative groups separately. The two groups differed significantly, $\chi^2(21) = 291.3$, $N = 6{,}000$, $p < 0.01$. The Positive group

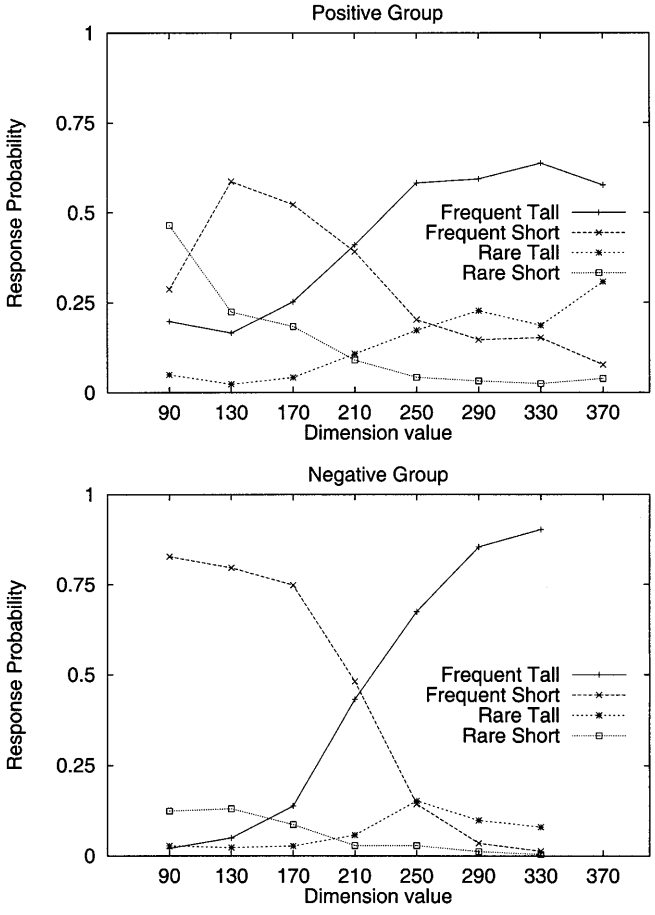

Fig. 4 The results of the last 100 trials of the training phase, in terms of the logical response categories. The categories are collapsed across dimensions: using the labels in Fig. 2, for the Positive group the tall frequent category (FT) is category B, the short frequent category (FS) is category C, the tall rare category (RT) is D on dimension 2 and A on dimension 1, while the short rare category (RS) is A on dimension 1 and D on dimension 2. For the Negative groups, the tall frequent category (FT) is D on dimension 2 and A on dimension 1, the short frequent category (FS) is A on dimension 1 and D on dimension 2, the tall rare category (RT) is category B, and the short rare category (RS) is category C. The absence of data in the final interval of the Negative group is due to the random ordering of stimuli during training

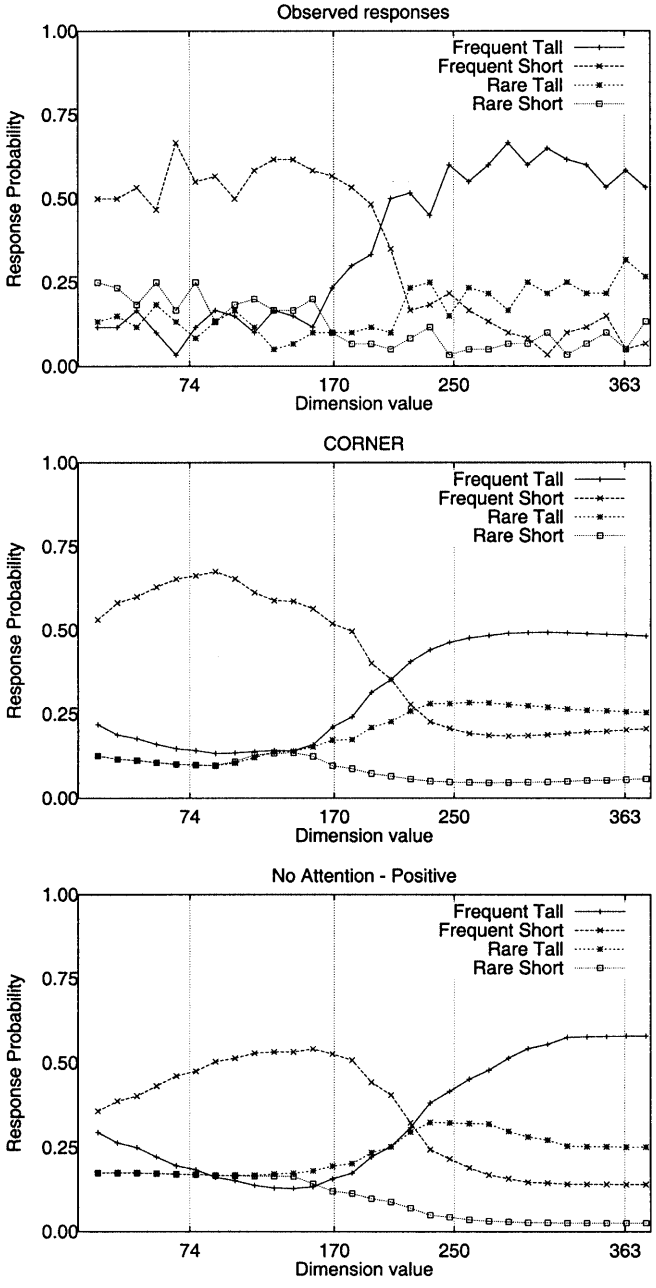

Fig. 5 The results of the transfer phase for the "Positive" training condition. *Top panel* The observed response frequencies. *Middle panel* The predicted response probabilities of CORNER. *Bottom panel* The predicted responses of CORNER without attention shifting or attention normalization. The categories are collapsed across dimensions: using the labels in Fig. 2, the tall frequent category (FT) is category B, the short frequent category (FS) is category C, the tall rare category (RT) is D on dimension 2 and A on dimension 1, while the short rare category (RS) is A on dimension 1 and D on dimension 2. *Vertical lines* indicate (from *left* to *right*) the lower limit of training stimuli, the mean of the short categories, the mean of the tall categories, and the upper limit of training stimuli

differentiated the two rare alternative from each other better than the Negative group, and responded with the rare labels more than the Negative group. The Negative group, in turn, used the frequent category labels more often than the Positive group, and also differentiated the two frequent categories better.

Transfer

The critical data for determining what was learned during training are the transfer trials. Again, responses were placed into logical categories, forming four response types. Responses to the transfer stimuli were aggregated over subjects within each group to form observed response probabilities, shown for the Positive group in Fig. 5 and for the Negative group in Fig. 6. Each subject contributed two independent responses (one for each physical dimension) at each of the 29 stimulus locations.

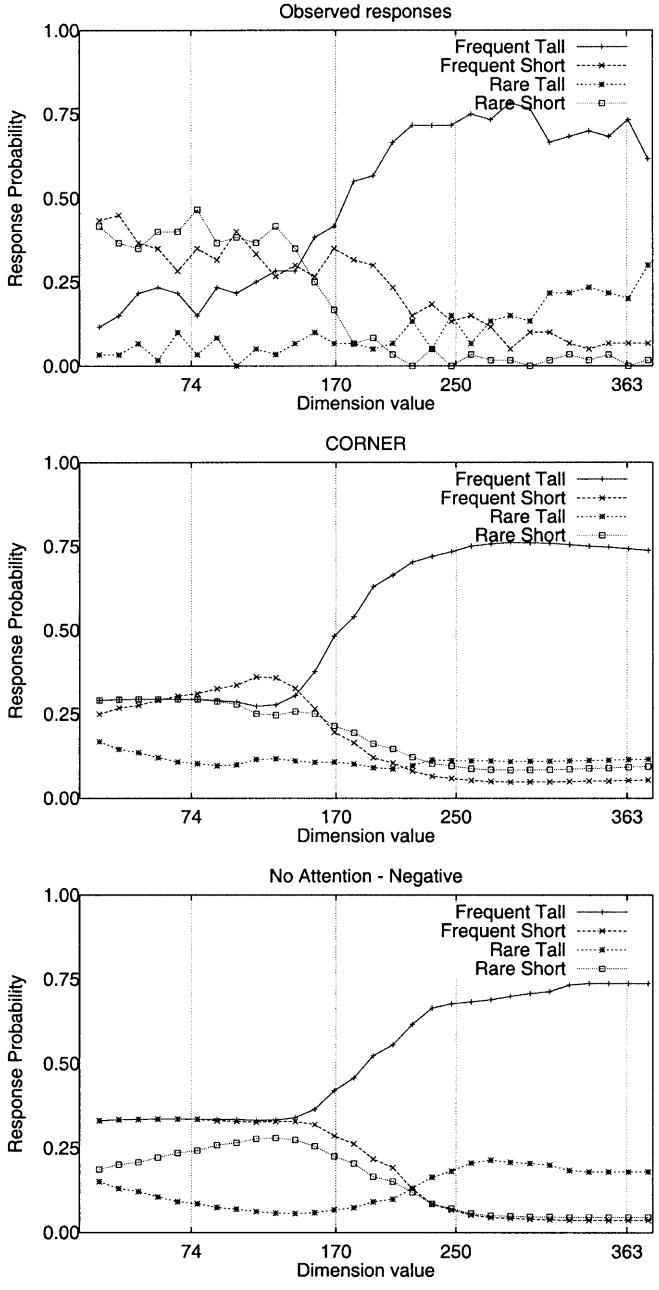

Fig. 6 The results of the transfer phase for the "Negative" training condition. *Top panel* The observed response, frequencies. *Middle panel* The predicted response probabilities of CORNER. *Bottom panel* The predicted responses of CORNER without attention shifting or attention normalization. The categories are collapsed across dimensions: using the labels in Fig. 2, the tall frequent category (FT) is D on dimension 2 and A on dimension 1, the short frequent category (FS) is A on dimension 1 and D on dimension 2, the tall rare category (RT) is category B, and the short rare category (RS) is category C. *Vertical lines* indicate (from *left* to *right*) the lower limit of training stimuli, the mean of the short categories, the mean of the tall categories, and the upper limit of training stimuli

Subjects in the Positive group chose one of the frequent categories more than any other category at all stimulus levels, preferring FT for all tall stimuli and FR for short stimuli. The peak response frequencies were located at more extreme stimulus values than were the category means. Throughout the range in which a particular frequent category was preferred, the corresponding rare category was also chosen more often than either the other rare or frequent category. The ratio of the frequent to rare choice was nearly 3:1, which was the presentation ratio. For the other two categories (for example, the short rare and frequent categories considered above the tall mean), the observed ratio of responses was closer to 1:1.

The situation is quite different for the Negative group. Subjects responded with the FT category most often for all stimuli situated at the tall end of the scale, with responses of the RT category generally increasing as the stimulus values become extreme. The probability of responding with label FT was also greater than the probability of responding with FS, even at the mean for the short categories. Subjects appeared to make little distinction between the FS and RS categories, making each response about equally often, with rare choices sometimes exceeding frequent as the stimulus value decreases. The ratio of responses between the FT and RT categories, in contrast, is greater than 3:1 throughout much of the stimulus range. It appears, therefore, that subjects associated the tall stimulus values with their frequent category labels, and learned little else about the stimulus values.

The results are mixed with respect to base-rate effects. Base-rate overuse appears near both means in the Positive group, and near the taller mean in the Negative group. Base-rate neglect appears to be the rule with extreme stimulus values in the Positive group and all short values in the Negative group, as rare categories are chosen relatively more than their posterior probabilities. Peak response rates are shifted away from the true category means. All trends in the data were continued through the extrapolation range, although this range was not very great.

Theoretical analysis

CORNER was fit to the transfer data using a mixture of simulated annealing and hill-climbing methods, with model fit calculated using the likelihood (G^2) statistic to relate predicted response frequencies to observed frequencies. The data have 174 degrees of freedom, as there are 29 stimulus values by three independent response levels (the fourth is fixed by the marginals) for each of the two training conditions. Each subject contributed 58 responses, which are taken to be independent; the $29 \times 4 \times 2$ contingency table therefore has an $N = 3{,}480$.

The effect of thermometer coding is to make tall stimulus values activate more units than short values and, thus, make more weights available to be changed by learning. Consider a stimulus with values (200, 50). Dimension One (with value 200) will activate four times as many value nodes as Dimension Two, and the weights connecting each active node to the category nodes will be adjusted equally. This means that participants will

associate Dimension One with the correct response more strongly than Dimension Two, and, in this way, categories that have one tall dimension and one short dimension will be identified by their tall dimension. Categories with no tall dimension (C in Fig. 2) will be at a disadvantage. Combined with attention normalization, this provides the basis for the distinction between the Positive and Negative training conditions.

The thermometer encoding reflects the extra attention that subjects devote to the taller stimulus values. However, attention is a limited resource; and attention normalization ensures that each active value node receiving an equal proportion of the available attention. When a stimulus is tall on both dimensions (symmetric), there are many active nodes, but attention to each node is low. When a stimulus is short on both dimensions (symmetric), there are few active nodes, attention to each node is high. When one stimulus dimension is tall and the other dimension is short (asymmetric), attention to each node is intermediate. Because attention and activation are multiplied, if attention to a node is low, the contribution of the node to the output nodes will be less. The learning of the long-term association weight connecting a value node to an output node depends on this contribution, and so a weight from a node that receives less attention changes less rapidly. The model assumes participants can attend differentially to the various elements of a stimulus within a single trial. Asymmetric representations combined with attention normalization, however, also produce differential attention for tall, short, and asymmetric stimuli across trials.

Thus, thermometer coding causes tall and short values to have unequal effects. Attention normalization makes symmetric and asymmetric stimuli different from each other, and attention shifting can amplify both of these asymmetries. For this reason, the effect of attention shifting per se can be difficult to visualize. The results of this experiment were analyzed by fitting versions of CORNER that implemented either attention normalization, error-driven attention shifts, both or neither. The relative success of these models can be used to determine the relative contributions of each of CORNER's theoretical constructs.

Table 1 contains the parameters of the best fitting models. While the no-attention model describes the responses fairly well, both attention normalization and attention shifting result in statistically better fits to the data [$\chi^2(1) = 73$ and $\chi^2(1) = 23$, $p < 0.01$ respectively]. The full CORNER model fits significantly better than either of these two models [$\chi^2(1) = 14$, $\chi^2(1) = 36$, $p < 0.01$ for normalization and shifting, respectively].

The middle panels of Figs. 5 and 6 show the predicted response probabilities of CORNER. The model does a good job of capturing most aspects of subjects' average responses. Compared to a null model of homogeneity of responses over all stimulus magnitudes, CORNER with seven free parameters accounts for approximately 98% of the variability in the data [$G^2(3) = 13016$, $R^2(\mathrm{U}) = 0.983$, $N = 3{,}480$].

Table 1 The best fitting parameters of the CORNER models. The parameters are explained in the text. Parameter values in bold are fixed (G^2 likelihood ratio statistic, *df* degrees of freedom, *p* probability of the observed response frequencies given the model)

Parameter	CORNER	No shifting	No normalization	No attention
ϕ	0.44	0.72	0.23	0.26
β	2.77	2.01	1.79	3.98
λ	1.32	10.0	0.06	0.034
λ_α	0.04	**0**	0.185	**0**
η	1.66	1.18	∞	∞
G^2(df)	183 (172)	197 (172)	233 (172)	256 (173)
p	0.25	0.09	<0.01	<0.01

As discussed above, both the asymmetric stimulus representation and a limited attention capacity are partly responsible for the model's good fit. The addition of attention-shifting increases the model's fit, so that all three constructs (thermometer encoding, attention normalization and attention shifting) contribute to CORNER's success. Attention shifting serves to emphasize asymmetries in the association weights that occur during training. An example will help make this clear.

Consider the Negative training condition, where the two frequent categories each contain one short and one tall stimulus dimension. At the onset of training, association weights are all zero. When a frequent category stimulus is presented, it creates an asymmetric distribution of activation between the two dimensions, activating a moderate number of nodes. Because attention is normalized, the amount of attention the model can allocate to each value node is moderate – less than when the values on both dimensions are short, but greater than when both dimensions are tall. Because association weights are uniform, there is no role yet for attention shifting. When the weights are adjusted, they increase moderately for the correct category and decreasing moderately for the incorrect category, reflecting the moderate activation of the nodes.

The asymmetric thermometer encoding, along with attention normalization, ensure that the distribution of association weights does not remain uniform for long. Stimuli that are short on both dimensions produce larger activations, due to normalization. The model consequently makes larger changes in the association weights for the activated nodes of short stimuli than for the same nodes when activated by tall stimuli. Input nodes that code for short stimulus values are always activated, and thus their weights have more opportunities to be updated. Early training constructs a gradient of association weights, as the frequent categories begin to be learned.

With a gradient of association weights, attention shifting can begin to play a role. The frequent categories are initially associated primarily with the tall value nodes of their tall dimension. When a rare stimulus is presented – for example, one that is short on both dimensions – the incorrect (frequent) category representations are activated. Early learning leaves the taller value nodes with the strongest association weights to

these incorrect categories. The learner can will shift attention away from the taller value nodes toward the shorter value nodes to reduce categorization errors. When activation and error are recomputed, the association weights from the shorter value nodes will change most, because attention to them was highest. The stronger association between short value nodes and the rare short category is also partly constructed by attention normalization, since the rare short category, with its two short stimulus dimensions, will result in maximum attention to the few active nodes. Attention shifting and attention normalization work together to distort the learner's representation of the rare category systematically. In the Positive condition, distortion is reduced because the common categories are harder to learn, due to the asymmetry of the thermometer encoding. A similar distortion applies to the frequent categories, but, because they are learned first, the effect is less.

The distortion effects produced by attention normalization and attention shifting can be seen by comparing the middle and bottom panels of Figs. 5 and 6, which show the full CORNER's predictions, and those of CORNER without attention, for the Positive and Negative conditions respectively.

Without attention shifting or normalization, the only asymmetry in processing comes from the asymmetric representation of the stimulus. This asymmetry results in differential treatment of tall and short stimulus values, because the total activation in the input layer is different, yielding differences in association weights between stimuli of different magnitudes. However, the differences are small, because there is no redistribution of attention to magnify them. The base-rate information, which is not sensitive to the stimulus values, dominates the predicted response probabilities. For example, the predicted frequency of FT responses is equal to that of FS responses for all short stimuli in the Negative condition, and the predicted frequency of FT responses is approximately constant for all tall stimuli in the Positive condition. CORNER, in contrast, learns more about each stimulus value, and therefore more about the rare categories. Notice that CORNER's predicted responses are not dominated by the relative base rates; in the Positive condition, for example, the two frequent categories are more distinct for CORNER than for the model without attention.

Despite the model's rough approximation to the psychological representation of the stimuli, the role of error-driven attention shifting is clearly demonstrated. The single attention learning parameter increases the fit of the model to both data sets to a significant extent, and only the complete CORNER model has non-significant lack of fit, with $p = 0.25$ (see Table 1).

General discussion

CORNER begins to fuse the concept of attention shifting with those concepts that were contained in earlier category-learning models, such as ALCOVE (Kruschke, 1992). CORNER differs from ALCOVE in two important ways. The representation of continuous dimensions used in CORNER has obvious limitations. Unlike ALCOVE, it does not include a representation of each stimulus configuration, but only a representation of each stimulus component. As a result the model cannot learn non-linear mappings, for the same reason that a simple perception cannot learn XOR. CORNER could be extended to include nodes whose activations are determined jointly by all stimulus dimensions, such as the configural nodes of the Configural-Cue model of Gluck and Bower (1988), or exemplar nodes such as those in ALCOVE. To add a complete multidimensional representation of each stimulus would require a mechanism for keeping attention dimensional (as in ALCOVE), while allowing it to shift within and between dimensions (as in CORNER). The current model makes this distinction implicitly, by applying more attention to whichever stimulus dimension has the largest value. These exemplar nodes could mediate arbitrary non-linear mappings between the input and category nodes, but attention would remain tied to stimulus dimensions and values.

Attention, as well as activation, is allocated differently during categorization by CORNER and ALCOVE. CORNER takes ADIT's notion of attention to discrete features and extends it to continuous features. However, ADIT's features are also dimensions, in the sense that each dimension only has two values (zero and one), and requires only one node to represent the dimension. ALCOVE also has a system of dimensional attention, a concept that is vital in explaining a variety of effects in category learning (Kruschke, 1992, 1993). There is an important difference between CORNER and ALCOVE's attention modification. In CORNER, attention is rapidly re-allocated on every trial in response to error. In ALCOVE, however, error serves to modify attention more slowly, but the modifications are long lasting – in other words, CORNER *shifts* dimensional attention but ALCOVE *learns* dimensional attention. Attention learning and attention shifting can sometimes produce similar results (Lewandowsky, Kalish, Griffiths, & Phang, 1999), but the two are conceptually distinct.

CORNER is clearly not a complete theory of category learning. Rapid shifts in attention to stimulus features, however, are evidently important in explaining how people learn to categorize stimuli with continuous-valued dimensions. CORNER's inability to learn non-linearly discriminable categories could be easily remedied.

Attention shifting and attention learning could be manifestations of a single process (Kruschke, 1999; Kruschke & Johansen, 1999). Kruschke has suggested that attention learning is the memorization of attention shifts – a form of pro-active shift that reduces expected error (Mackintosh, 1975). Such internalization could be done through an adaptive 'gate' (Jacobs, Jordan, Nowlan, & Hinton, 1991) that mixes the activations of individual stimulus dimensions. The gate could contain

exemplar nodes, of the sort necessary to form non-linear mappings. Such a system could potentially divert attention differentially to individual sub-spaces, while maintaining a multi-dimensional representation to enable non-linear mappings. Whether attention shifting and attention learning prove to be related concepts or not, it is clear that the rapid shifting of attention to stimulus values plays an important role in how people learn to categorize both continuous and discretely dimensioned stimuli.

Acknowledgements Thanks to Abbey Clawson, Beth Okeon, Laura Owen, Debbie Reas, and Daniel Vote for assistance in administering the experiments. The author's World Wide Web sites are http://www.psy.uwa.edu.au/user/kalish and http://www.indiana.edu/~kruschke. This research was partially supported by NIMH training grant T32 MH19879-02 and ARC grant A79941108 to Michael Kalish and by NIMH FIRST award 1-R29-MH51572-01 to John Kruschke.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review, 84,* 413–451.

Ashby, F. G., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

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

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93,* 154–179.

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

Estes, W. K. (1994). *Classification and Cognition*. Oxford, UK: Oxford University Press.

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

Green, D. M., & Swets, J. A. (1967). *Signal detection theory and psychophysics*. New York: Wiley.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 344–354.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1–16.

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

Kruschke, J. K. (1993). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: the psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego, CA: Academic Press.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 3–26.

Kruschke, J. K. (1999). Toward a unified model of category learning. Available on-line at http://www.indiana.edu/~kruschke/tumaal.html.

Kruschke, J. K., & Johansen, M. K. (in press). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Lewandowsky, S., Kalish, M. L., Griffiths, T. L., & Phang, J. (1999). Knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 103–189). Wiley, New York.

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

Maddox, W. T. (1995). Base-rate effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 21, 288–301.

Marley, T. (1992). Relations between the grt and gcm. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 299–334). Hillsdale, NJ: LEA.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General,* 117, 68–85.

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

Nosofsky, R. M. (1988). On exemplar-based exemplar representations: reply to Ennis (1988). *Journal of Experimental Psychology: General,* 117, 412–414.

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 322.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review, 3,* 222–226.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.