THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 1997, *50A* (3), 607–630

Generalization in Human Category Learning: A Connectionist Account of Differences in Gradient after Discriminative and Non-discriminative Training

A.J. Wills and I.P.L. McLaren
University of Cambridge, Cambridge, U.K.

Two experiments are reported that investigate the difference in gradient of generalization observed between one-category (non-discriminative) and two-category (discriminative) training. Extrapolating from the results of a number of animal learning studies, it was predicted that the gradient should be steeper under discriminative training. The first experiment confirms this basic prediction for the stimuli used, which were novel, prototype-structured, and constructed from 12 symbols positioned on a grid. An explanation for the effect, based on the Rescorla–Wagner theory of Pavlovian conditioning (Rescorla & Wagner, 1972), is that under non-discriminative training "incidental stimuli" have significant control over responding, whereas under discriminative training they do not. Incidental stimuli are those aspects of the stimulus, or the surrounding context, that are not differentially reinforced under discriminative training. This explanation leads to the prediction that a comparable effect of blocked versus intermixed discriminative training should also be found. This prediction is disconfirmed by the second experiment. An alternative model, still based on the Rescorla–Wagner theory but with the addition of a decision mechanism comprising a threshold unit and a competitive network system, is proposed, and its ability to predict both the choice probabilities and the pattern of response times found is evaluated via simulation.

The concept of a generalization gradient is perhaps one of the most pervasive in psychology. For a very wide range of situations, and indeed species, it can be concluded that if a response is trained to a stimulus, then the likelihood of another stimulus also evoking that response is a function of the similarity between them. When testing rats, rabbits, or pigeons, reliable generalization gradients on a variety of simple physical dimensions have been found; examples include the wavelength of a light (Guttman & Kalish, 1956), the frequency of a tone (Moore, 1972), and the size of a circle (Grice and Saltz, 1950). Work with pigeons demonstrates that orderly generalization gradients can also be

Requests for reprints should be sent to A.J. Wills, The Psychological Laboratory, University of Cambridge, Downing St., Cambridge, CB2 3EB. Email: ajw43@cus.cam.ac.uk

This research was supported by a grant from the BBSRC to I.P.L. McLaren. The authors would like to thank N.J. Mackintosh for his help in the early stages of this project, and J.K. Kruschke and an anonymous reviewer for their comments on earlier versions of this manuscript.

observed with rather more complex stimuli: Brunswik faces, for example (Huber & Lenz, 1993), or shapes varying in colour, number, and form (Jitsumori, 1993), or seed-like stimuli (Lea, Lohmann, & Ryan, 1993). Shepard has demonstrated that, within a psychological space determined by the mistakes made in an identification learning task, human subjects show an orderly (exponential) generalization gradient, and that this holds for stimuli as diverse as circles of various sizes, vowel phonemes, and morse code signals (Shepard, 1987). Humans also show generalization gradients in tasks procedurally more similar to those used in the animal work (e.g. Buss, 1950).

One of the empirical facts about generalization gradients in animals is that they are sharpened by discrimination training. The robustness of this effect is well established within animal learning research (e.g. Hanson, 1959; Jenkins & Harrison, 1960; Newman & Baron, 1965), and the generally accepted explanation is that discriminative training neutralizes the effect of "incidental stimuli" (see Mackintosh, 1974 for a discussion). Incidental stimuli are aspects of the stimulus, or of the surrounding context, that are not differentially reinforced under discriminative training. For example, in the Newman–Baron study, the stimulus was a single vertical white line on a green background. Pigeons were reinforced for pecking it and then tested on lines of varying orientation (from upright to 45° either side), also on a green background. They showed a basically flat generalization gradient unless non-reinforced presentations of the background alone were also included in training. The explanation offered for this result is that in the absence of differential reinforcement the green background is as good a predictor of food as the line and therefore has significant control over responding. As the background is present in all test examples, a shallow generalization gradient is seen. Associating the background alone with the non-occurrence of food reduces its predictive value and hence increases the steepness of the generalization gradient.

The Rescorla–Wagner model of Pavlovian conditioning (Rescorla & Wagner, 1972) may be employed to give a more precise statement of this general line of reasoning. The model states that, on any given trial, the change in strength of the association between a conditioned stimulus (CS) and the unconditioned stimulus (UCS) is:

$$\Delta w = \alpha \cdot \beta \cdot (\lambda - \sum^{N} w_i) \qquad 1$$

where λ is the asymptote of learning for the US, and Σw_i is the sum of the associative strengths for all N CS present on that trial. α and β control the rate of learning; β is assumed to be determined by the salience of the US and α by the salience of the relevant CS. To see how Equation 1 predicts a steeper gradient of generalization under discriminative training, consider a simple case where training involves just two conditioned stimuli: the target stimulus (CS_T) and an incidental stimulus (CS_C) of equal salience. In non-discriminative training, CS_T and CS_C occur together and are reinforced; they therefore gain associative strength as a function of their salience and, at the limit, this will be equal to $\lambda/2$. In discriminative training, CS_T and CS_C also occur together in the presence of reinforcement, but, in addition, CS_C occurs alone in the absence of reinforcement. For non-reinforced trials, λ will be zero, so CS_C will lose associative strength on these trials. This will result in CS_T's associative strength rising more quickly than CS_C's,

and the difference between the two will be further compounded by Σw_i approaching λ on reinforced trials. Learning stops when CS_T has an associative strength of λ and CS_C has an associative strength of zero. Hence the gradient of generalization in the discrimination condition will be steeper as CS_C, presumably constant on test, has little control over responding. Therefore, changes in CS_T will have a greater effect on behaviour.

The purpose of the present paper is to investigate whether a similar generalization gradient difference between discriminative and non-discriminative training can be seen in human category learning and, if so, whether an explanation similar to the one given above is appropriate. The acquisition of two or more novel categories has been the subject of many previous studies, and some of these provide data that may be used to assess generalization gradients. Examples include the study of prototype effects (e.g. Posner & Keele, 1968), of probabilistic cue learning (e.g. Estes, Burke, Atkinson, & Frankmann, 1957), and of the diagnosis of imaginary diseases (e.g. Estes, 1986). Some of the work on exemplar theories of categorization (e.g. Nosofksy, 1986, 1991) is also applicable. However, none of these studies has the non-discriminative training condition needed for comparison with discriminative training. The studies reported in this paper provide the appropriate control groups.

It seems reasonable in principle to extend the Rescorla–Wagner-based explanation to human category learning. Gluck and Bower (1988) have previously demonstrated that it can be used (with a slight modification detailed below) to predict aspects of human subjects' performance in a simulated two-disease medical diagnosis paradigm. This situation is not radically different from a standard categorization experiment. Furthermore, as a number of authors have noted, Equation 1 is in many ways equivalent to the delta rule, an error-correcting algorithm widely used in connectionist models (e.g. McClelland & Rumelhart, 1985; Rumelhart, Hinton, & Williams, 1986). Such models have had some success in explaining certain aspects of human classificatory behaviour.

In the Gluck–Bower paper, the two diseases are represented by $+\lambda$ and $-\lambda$ rather than the $+\lambda$ and 0 used to represent reinforcement and its absence in the Rescorla–Wagner model. This modification allows both diseases to cause learning on the first trial, even when the associative strengths start at zero. In this form, the Gluck–Bower model can potentially be applied to any two-choice category learning experiment by representing each significant aspect of the two categories with a different CS. If, in addition, the subject were representing some components of the stimulus incidental to the current discrimination, then one might expect to see an effect analogous to the difference between discriminative and non-discriminative training. In other words, presenting the subject with examples from just one of the two categories should result in a shallower gradient of generalization around the category locus than if training involves both categories.

This may be clearer with a specific example. Consider the simplest case of three equally salient stimulus components: one a component perfectly predictive of one category, another similarly predictive of the other, and the third with no predictive value but which always appears in compound with one of the other two. If just one category is presented, then the perfectly predictive component and the incidental one gain equal associative strength. If both categories are presented intermixed, then, on any trial, the incidental component is equally likely to be paired with either of the categories. This will result in opposing changes in associative strength, which will cancel. Learning stops with

the two perfect predictors having equal associative strength of the opposite sign and the incidental stimulus having an associative strength of zero. If a generalization test is performed that systematically varies the similarity of the perfect predictor component to the one used in non-discriminative training whilst the similarity of the incidental component to the trained one stays constant or varies unsystematically, then the generalization gradient should be steeper in the two-category (discriminative) than the one-category (non-discriminative) training condition.

Ideally, research on this issue would use categories that allowed stimulus factors to be predictive or non-predictive. One stimulus structure that allows this can be found in studies of what are often described as polymorphous concepts (Dennis, Hampton, & Lea, 1973). The name is intended to cover any collection of stimuli whose category membership is defined by an *m*-out-of-*n* rule; an example would be "at least two of symmetric, black, and composed of circles". In animal studies, at least, orderly generalization gradients have been observed around such categories (Huber & Lenz, 1993; Jitsumori, 1993; Lea, Lohmann, & Ryan, 1993). Furthermore, their representation in the models discussed above is unproblematic: each of the *n* may be considered as one partially predictive stimulus component. Hence it seemed reasonable to study the effect of type of training (discriminative versus non-discriminative) on the generalization gradient observed around polymorphous categories. This was the purpose of the first experiment.

EXPERIMENT 1

Each of the stimuli used in this experiment was an array of spatially separate symbols (elements) whose position in that array conveyed no information. Stimuli consisting of arrays of separable elements have been used extensively in the study of categorization with humans (e.g. Medin & Schaffer, 1978; Regehr & Brooks, 1995; Shepard, Hovland, & Jenkins, 1961), although the ones used here differ in that the underlying structure is basically polymorphous and the position of elements is unimportant.

Comparing two-category learning with one-category learning raises two methodological problems. The first is how meaningful category membership information may be provided in a one-category (non-discriminative) situation. The typical "guess and correct" method of two-category (discriminative) studies seems inappropriate, as the answer will always be the same. Here this problem is resolved by presenting the category label alongside the stimulus in both the one- and the two-category conditions. The second problem is to determine the appropriate number of training trials in the non-discriminative condition—one can either control for the total number of stimuli seen or the number for a particular category. A discussion of the relative merits of these two systems is avoided here by including both.

Method

Subject and Apparatus

The subjects were 36 students from Cambridge University, mostly undergraduates, who were paid for their participation. They were tested individually in a quiet experimental cubicle. The experiment was presented on a colour monitor (Acorn AKF60, 27 × 20 cm) in a medium-resolution

(800 × 600 pixels) 16-colour mode, and responses were made on a standard computer keyboard, both of which were connected to an Acorn RiscPC 600 microcomputer. Subjects sat about 1 metre from the screen, which was approximately at eye level.

Stimuli

All stimuli were composed of 12 different small pictures (elements) placed inside a thin rectangular outline measuring 4.5 × 3.5 cm. Each element occupied a different position on an invisible grid 4 items across and 3 down, this position being determined randomly. The 12 elements of a stimulus were drawn randomly for each subject from a pool of 24, and for each subject this pool was a new, randomly selected sub-set of the elements shown in Figure 1.

The way in which elements were drawn from this sub-set to create stimuli depended on whether they were to be used for training or for test. At the beginning of the experiment, 12 elements were randomly designated as being generally predictive of one category ("A"), and the remaining 12 as generally predictive of the other ("B"). This allocation was done even if only one category would actually be shown in training. Training examples were created by starting with the 12 appropriate

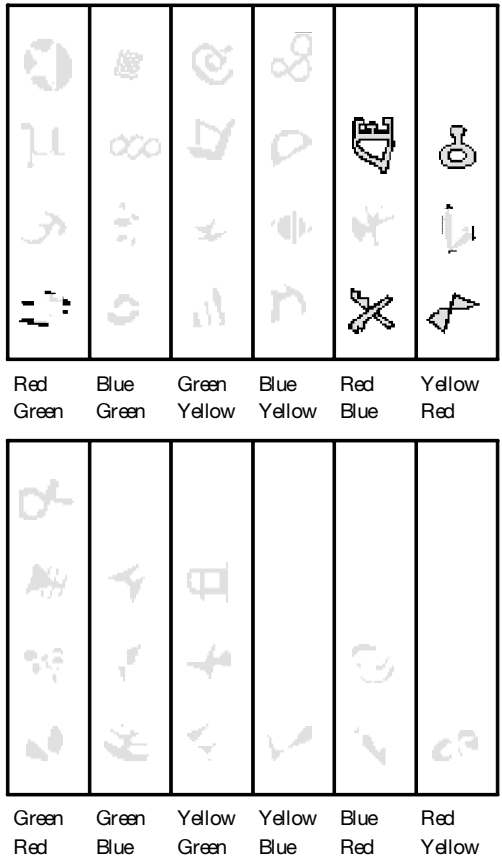

FIG. 1. The 36 small pictures used to construct the stimuli in Experiments 1 and 2. Column labels denote the outline and fill colours, outline uppermost.

predictive elements and giving each a small independent chance of being replaced by an element from the other category. Replacement elements were chosen at random, with the constraint that no element appear twice in the same stimulus. In terms of the number of predictive elements, the stimuli can be characterized by the binomial distribution ($p = 0.9$, $N = 12$) shown in Figure 2. (This category structure is similar to that produced by an "at least 8 out of 12" rule and is directly analogous to the form used in McClelland & Rumelhart's, 1985, modelling paper.) Test exemplars were created from a specified number of generally predictive elements from each of the two categories, which always summed to 12. Apart from situations where the stimulus consisted entirely of predictive elements from one category, more than one combination of elements was possible. In these situations, the elements to be used were chosen randomly for each stimulus.

Procedure

Twelve subjects were allocated to each of three conditions. After some general instructions and the presentation of an example labelled stimulus, subjects in the discriminative training (A30/B30) condition were presented with 30 examples of Category A and 30 examples of Category B sequentially and in a random order. Each example was presented for 5 sec in the centre of the monitor and was accompanied by its label, which was presented as a large sans-serif capital A or B in an outline rectangle (4.5 × 3.5 cm) immediately to the right of the stimulus; 2 sec of a plain mid-grey mask in the stimulus and label rectangles preceded the following example. Subjects were not required to respond in any way; they were simply asked to concentrate, as they would later be asked to classify new, unlabelled examples. In the non-discriminative training conditions, either 30 examples of Category A were presented (condition A30), or 60 examples (condition A60). Again, the stimuli were accompanied by a label (always A in these conditions), and subjects were told they would later have to classify unlabelled examples. No examples of Category B were presented during training in the non-discriminative conditions.

The training phase was followed by a test phase, in which 130 test stimuli, 10 from each of the 13 positions in the sequence (12 A elements, no B elements) to (no A elements, 12 B elements), were presented in a random order. In the A30/B30 condition, subjects were asked to make a forced choice for each stimulus from the options "It is an A" and "It is a B". They were encouraged to respond as quickly as possible, but the machine would in fact wait indefinitely for a response. As soon as a response had been made, it was recorded, along with the time taken to make it, and the following stimulus was presented immediately. The test procedure was very similar in the A30 and A60 conditions; the option "It's a B" was replaced with the option "It's not an A".

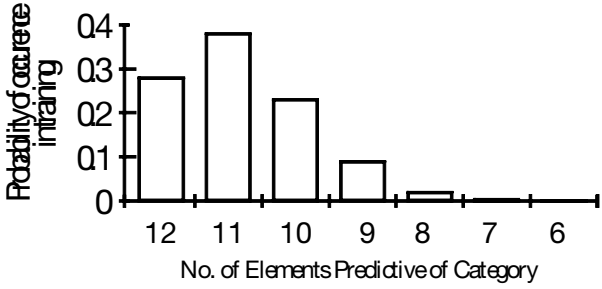

FIG. 2. The probability distribution for the stimuli used in Experiments 1 and 2. Stimuli always comprise 12 elements, but the number of those predictive of the category of which the stimulus is a member varies. The distribution is binomial ($p = 0.9$, $N = 12$).

The allocation of keys to responses was counter-balanced across subjects. Within each condition, half the subjects used the "X" key to respond, with their left hand, "It's an A", and the "." key to respond, with their right hand, "It's a B" or "It's not an A". The other half were given the reverse assignment.

Results

In case any subject had inadvertently reversed the key-to-key category assignment, subjects' performance at points 0 B elements and 12 B elements was assessed for being significantly below chance (binomial test, $N = 10$, $p = 0.05$). Any subject who was below chance at both points would have been replaced, but, in fact, no subject failed this test. The counter-balance sub-conditions did not differ significantly, so they were collapsed in all subsequent analyses, for which the significance level is taken to be .05.

As can be seen from Figure 3(a), all conditions show an orderly generalization gradient; as the number of elements generally predictive of Category B in a stimulus increases, the probability that subjects will call it an "A" decreases. A mixed analysis of variance (ANOVA), with one within-subject variable (B elements, 13 levels) and one between-subjects variable (experimental condition, 3 levels) showed that this effect was significant, $F(12, 396) = 150$, as was the difference between conditions, $F(2, 33) = 11$. A post-hoc Tukey (HSD) test revealed that the latter was due to condition A30/B30 differing significantly from both condition A30 and condition A60. Conditions A30 and A60 were not significantly different. There was also a significant Condition × Number of B Elements interaction, $F(24, 396) = 2.9$.

From inspection of Figure 3(a), it would appear that the gradient in the A30/B30 condition is steeper than in the A30 or A60 conditions in the range 0–6 B elements and shallower in the range 6–12 B elements. Linear regression was used to assess this, but first the data were checked for an overall significant linear component. This was found in all conditions and sub-ranges: the relevant F-ratios are $F(1, 82) = 97$, 20, and 39, for

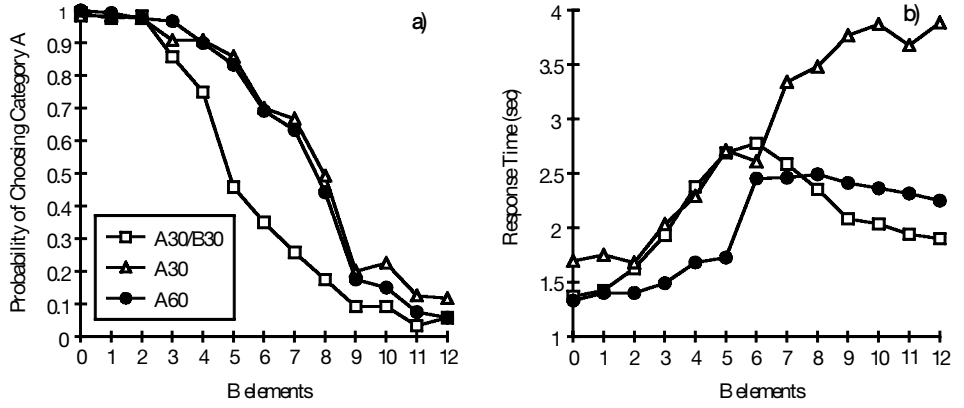

FIG. 3. Probability of subjects choosing category A and the mean time taken to do so after discriminative (A30/B30) and non-discriminative (A30 and A60) training in Experiment 1.

conditions A30/B30, A30, and A60, respectively, in the range 0–6 B elements, and $F(1, 82) = 24$, 52, and 90 in the range 6–12 B elements.

To assess the difference in gradients, separate regression lines were obtained for each subject in each of the two ranges. It was found that, in the range 0–6 B elements the gradients of the lines in condition A30/B30, mean = −0.11, were significantly more negative than the gradients of the lines in the A30 condition, mean = −0.04, $t(22) = 3.8$, and the A60 condition, mean = −0.05, $t(22) = 3.7$. The gradients in the A30 condition did not differ from those in the A60 condition, $t(22) = 0.76$. In the range 6–12 B elements the A30/B30 gradient, mean = −0.05, was significantly less negative than the A30 gradient, mean = −0.11, $t(22) = 2.1$, and the A60 gradient, mean = −0.12, $t(22) = 3.0$. Again, gradients in A30 and A60 conditions did not differ, $t(22) = 0.25$.

Some care must be taken in interpreting response times from an experiment that had no specific time-out procedure, but, nevertheless, they provide useful additional information. As can be seen from Figure 3(b), response time was affected by the number of B elements a stimulus contains. Another mixed ANOVA revealed that this effect was significant, $F(12, 396) = 17$. Although the main effect of condition was not significant, $F(2, 33) = 2.6$, the Experimental Condition × Number of B Elements in the Stimulus interaction was, $F(24, 396) = 3.6$. Inspection of Figure 3(b) suggests that this interaction is due to the A30 and A60 conditions showing an increasing trend in response time (possibly levelling off in condition A60), whereas the A30/B30 condition shows an inverted-U trend. In order to provide increased sensitivity for detecting trends, the response time data were recoded by subtracting the individual subject mean response time from each of the data points provided by that subject. After this recoding, all three conditions showed a significant linear component, $F(1, 154) = 6.2$ for condition A30/B30, 125 for condition A30, and 66 for condition A60. However, both the pattern of means shown in Figure 3(b) and inspection of the residuals suggested that there was also a quadratic component in some of the conditions. Polynomial regression revealed that this was significant in conditions A30/B30, $t(154) = 6.4$, and A60, $t(154) = 3.0$, but not in condition A30, $t(154) = 0.71$.

Discussion

Three main conclusions may be drawn from this experiment. First, people can discriminate categories with a polymorphous structure, and training them to do so leads to an orderly gradient of generalization around the category loci. This conclusion is in line with much of the previous research on human category learning (e.g. Estes, 1986; Posner & Keele, 1968). Second, the gradient of generalization is sharper after discriminative than after non-discriminative training. This conclusion is unaffected by the way in which the number of stimuli presented in the non-discriminative condition is controlled for. At least for the subjects and number of discriminative training trials used here, one may either control for the total number of stimuli seen or for the number of examples of the appropriate category. Although the discriminative versus non-discriminative generalization gradient difference is well established with rats and pigeons, this would appear to be the first direct demonstration of it in humans. Finally, the pattern of response times are generally in line with existing data on human categorization. Without labouring the point,

if number of A elements is an approximate index of typicality, then the pattern is congruent with Rosch's work on semantic categories (Rosch, 1973). Data from the study of categorization response times from a decision-bound perspective is also of a similar form (cf. Ashby, Boynton, & Lee, 1994). The difference between response times in the A30 and A60 conditions may be due to differences in familiarity with particular stimulus components; this last point is developed in the modelling section of the paper, where the response time data are considered in greater detail.

What is not clear from this experiment is whether the central result—that is, the gradient of generalization being sharper under discriminative than under non-discriminative training—should be explained in the same way as the analogous result in rats and pigeons. A first objection might be that, given that no element the subject sees is equally predictive of the two categories, there is nothing that might be considered as an incidental stimulus, and therefore no difference between gradients for discriminative and non-discriminative training should be expected. However, to make this objection is to forget that the experimenter's definition of relevant aspects of the stimulus may not accord with the subject's. Any number of constant factors in the stimulus or its context may form part of the subject's category definition in the non-discriminative condition. Some examples might be the number of pictures, their arrangement in a 4×3 matrix, their enclosure in a rectangle, or the position of the stimulus on the screen. In the discriminative condition, it would become clear that these elements of the stimulus representation were incidental. At the level of intuition, readers must judge for themselves how likely it is that such factors do in fact form part of the stimulus representation in humans. However, in animal experimentation, the generalization gradient difference is seen even when there is only one unitary conditioned stimulus from the experimenter's point of view (e.g. a pure tone: Jenkins & Harrison, 1960). The incidental stimuli hypothesis permits that the critical non-differentially reinforced aspects of the situation may not be part of the experimenter's definition of the stimulus.

Although the incidental stimuli hypothesis can comfortably explain the results of this experiment, it is not the only credible explanation for the pattern of results found. One could argue that the different training procedures in the discriminative and non-discriminative conditions lead to different decision processes at test. After discriminative training, one might expect the decision to be relative ("Is this more like an A, or more like a B?"), whereas after non-discriminative traning it would be absolute ("Is this an A, or isn't it?"). The latter decision requires comparison to some minimum threshold of category membership and could lead to a shallower generalization gradient without the need to invoke incidental stimuli. To disentangle these two explanations, one needs a situation in which the incidental stimulus hypothesis would predict a gradient difference even though the underlying decision process was the same.

One effective manipulation is to include a discriminative training condition where all examples of one category are presented before any examples of the other category (blocked training) and compare it to the more standard intermixed training. If incidental stimuli are instrumental in causing the difference between discriminative and non-discriminative training, then they should also cause a difference between blocked and intermixed discriminative training. Recall that, under intermixed discriminative training, the Gluck–Bower model predicts that all associations from the incidental stimulus to a

category representation tend towards zero. At first sight it might appear that this is also true for blocked discriminative training, but it is not. The model is error-correcting and, as such, is sensitive to the order in which stimuli are presented.

This is most easily illustrated by a simplified situation involving two categories, two perfect predictors and an incidental stimulus, all of equal salience. Assuming that the first category presented is represented by $+\lambda$ (although the argument also works if you assume it is represented by $-\lambda$) it is clear that, in the first phase of blocked training, the associative strengths for the perfect predictor of that category and for the incidental stimulus should rise at the same rate. If learning reaches asymptote, then both strengths should equal $\lambda/2$. When, in the second phase, the other category (represented by $-\lambda$) is repeatedly presented, the associative strength for the other perfect predictor will tend towards $-\lambda$, as will the associative strength for the incidental stimulus. However, at the beginning of this phase, the associative strength for the incidental stimulus was positive, and for the perfect predictor it was zero. This means that the latter will end up significantly more negative, and, if learning in this phase is also asymptotic, they will be $-\lambda$ and $-\frac{3}{4}\lambda$, respectively. These resulting associative strengths differ in two ways from those achieved in the intermixed condition: The incidental stimulus is somewhat associated to the second category presented, and the perfect predictor for the first category presented is less associated to its category label than the perfect predictor for the second category is to its. As before, we assume that in a generalization test the similarity of the perfect predictor components to those seen in training is systematically varied, whereas similarity of the incidental component to the trained one stays constant or varies unsystematically. Under blocked training, the non-zero strength of the association between the incidental stimulus and the category node leads to a bias to respond in favour of the second category presented. The absence of this bias after intermixed training means that the model predicts that the gradient of generalization should be shallower under blocked discriminative training than under intermixed discriminative training.

Two important facts about the previous analysis need to be appreciated. First, the conclusions drawn are not dependent on the total number of stimulus elements, or on the perfect nature of the predictors, or on learning being asymptotic. Second, this model does not unavoidably predict a gradient difference. If there are no significant incidental stimuli, then the gradients would be equal, but one could not then appeal to the same incidental stimuli to explain the difference between discriminative and non-discriminative training.

It may be easier to appreciate this with the aid of a specific demonstration, and to this end the Gluck–Bower model was implemented in a form more directly interpretable from the perspective of the previous experiment. Briefly, 24 predictive stimulus component nodes were created, 12 for each category. As in the experiment, 12 predictive components were present in any category example, each having a 10% chance of being a component generally predictive of the other category. In addition 12 incidental components were always active. Each model received 30 different randomly selected examples of each of the two categories, either blocked or intermixed. Where training was blocked, all Category B examples were presented first. The learning rate parameter ($\alpha \cdot \beta$) was set to 0.001. This was chosen to show that the effect occurs in non-asymptotic conditions, but the specific value is not critical. Finally, both models were tested (associative strengths frozen) with 10 examples of each of the stimuli in the range (12 Category A components, no Category B

components) to (no Category A components, 12 Category B components). The activation of each representation was computed by summing the associative strengths to that representation for all stimulus components present on that trial. Following Gluck and Bower's original paper (Gluck & Bower, 1988), the resultant activation was transformed into a probability with the following equation:

$$P(A) = 1/(1 + e^{-kA}) \quad (2)$$

where $P(A)$ is an index of the probability of choosing category A, k is a scaling constant, and A is the resultant activation. The constant k was set to 12 for this demonstration.

The filled symbols in Figure 4 show the mean simulated probability of choosing Category A, taken over 20 simulation runs. The result of simulations employing no incidental stimulus components are also presented (hollow symbols) to illustrate that their removal can lead to a prediction that generalization gradients in the two discriminative conditions should be equal. However, their removal logically leads to an inability to explain the difference between discriminative and non-discriminative training in these terms.

EXPERIMENT 2

The second experiment was basically a replication of the first, with the addition of a blocked discriminative training condition. The basic prediction was that if a difference between discriminative and non-discriminative training was again observed and it was to be explained by the neutralization of incidental stimuli, then a blocked versus intermixed difference, in the same direction, should also be seen. As the two non-discriminative conditions of Experiment 1 did not differ significantly in their gradients of generalization, it seemed unnecessary to run both. We decided not to include the A30 condition.

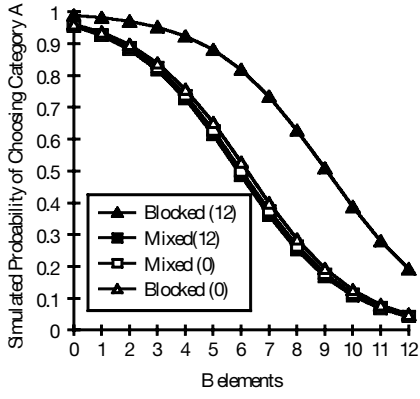

FIG. 4. Results from simulation of the effect of 0 and 12 incidental stimulus representations on generalization under intermixed and blocked discriminative training. Both mixed results are subsumed within one line as the differences between them are too small to be represented.

Method

The subjects were 67 adults, aged betwen 18 and 40, who were paid for their participation. None of the subjects from Experiment 1 was used. About half were tested using the apparatus described in Experiment 1, the rest on an AKF50 monitor and Acorn A5000 microcomputer in an adjacent cubicle. (This computer system is an older version of the AFK60/RiscPC600 system, and the program used runs identically on both machines without modification.) The stimuli were constructed in an identical manner to those used in Experiment 1. The procedure was identical to that in Experiment 1. In the new blocked (B30→A30) condition, all examples of B were presented before any examples of A.

Results

Key-to-category assignment was checked (in the same way as Experiment 1) for each subject. This time, some subjects were found to have reversed the assignment they had been given. Each subject who failed in this way was replaced by another person performing the same condition with the same key assignment, until there were 10 successful subjects in each condition's counter-balance sub-group (in total, 7 subjects were replaced). At this point, the counter-balance sub-conditions did not differ significantly from each other. The significance level for all analyses is .05.

As seen in Figure 5(a), all conditions showed an orderly generalization gradient; as the number of B elements increases, the tendency to respond "It's an A" decreases. A mixed ANOVA, with one between-subject variable (experimental condition, 3 levels) and one within-subject variable (B elements, 13 levels), showed that this effect was reliable, $F(12, 684) = 230$, as was the difference between conditions, $F(2, 57) = 31$. Post-hoc Tukey (HSD) tests revealed that this was due to the A60 condition being significantly different from the A30/B30 and B30→A30 conditions, which did not differ. The Number of B Elements × Experimental Condition interaction was also significant, $F(24, 684) = 6.2$.

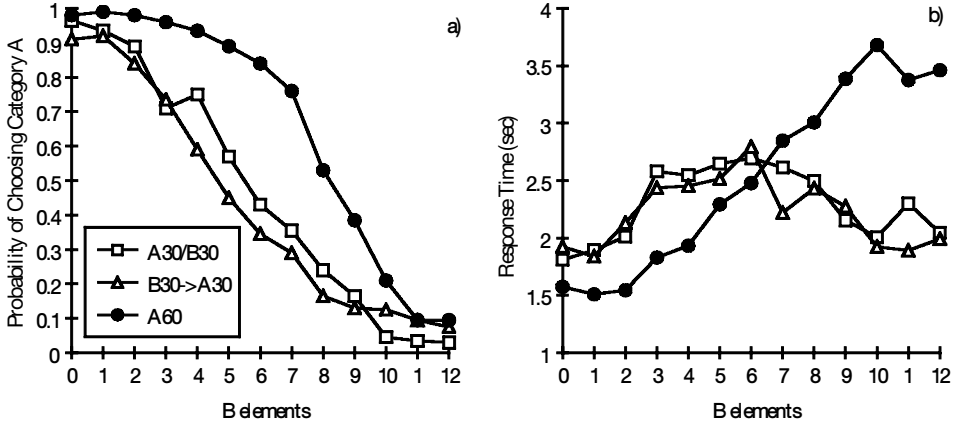

FIG. 5. Probability of subjects choosing category A (a) and the mean time taken to do so in seconds (b) after intermixed discriminative (A30/B30), blocked discriminative (B30→A30) and non-discriminative (A60) training in Experiment 2.

From inspection of Figure 5(a), it would appear that the gradients in the A30/B30 and B30→A30 conditions are steeper than in the A60 condition in the range 0–6 B elements but shallower in the range 6–12 B elements. Before assessing this, both ranges in all conditions were checked for an overall significant linear component, in all cases the relevant F-ratios were $F(1, 138) = 25$, 86, and 180 for conditions A60, A30/B30 and B30→A30, respectively, in the range 0–6 B elements, and $F(1, 138) = 140$, 84, and 41 in the range 6–12 B elements.

Once again, differences in gradient were assessed by obtaining separate regression lines for each subject and for each of the two ranges. It was found that, in the range 0–6 B elements, the gradients of the lines in the A60 condition, mean = −0.02, were significantly less negative than the gradients of the lines in the A30/B30 condition, mean = −0.09, $t(38) = 5.1$, and the B30→A30 condition, mean = −0.10, $t(38) > 10$. The A30/B30 condition did not differ from the B30→A30 condition significantly, $t(38) = 1.1$. In the range 6–12 B elements, the A60 gradients, mean = −0.14, were significantly more negative than the A30/B30 gradients, mean = −0.07, $t(38) = 4.30$, and the B30→A30 gradients, mean = −0.04, $t(38) > 10$. In this segment, the difference between the B30→A30 gradients and the A30/B30 gradients just reached significance, $t(38) = 2.2$, with the B30→A30 gradients being less negative.

The response times provided useful secondary information. As can be seen from Figure 5(b), they appear to show a similar pattern to Experiment 1. A mixed ANOVA revealed that the number of B elements had a reliable effect on response time, $F(12, 684) = 8.7$, and that this effect differed according to condition, as evidenced by a significant B Elements × Condition interaction, $F(24, 684) = 6.8$. To assess any possible trends, the response times at each test point were recoded for each subject by subtracting their personal overall mean response time from their mean response time at each test point (the same tranformation as used in Experiment 1). After this transform, the data in the A60 condition showed a reliable linear trend, $F(1, 258) = 160$. No reliable linear component was detected in the A30/B30 or B30→A30 conditions, $F(1, 258) = 0.19$ and $F(1, 258) = 0.65$, respectively. Second-order polynomial regression led to regression lines that described a significant component of response times in conditions A30/B30, $F(2, 257) = 14$, B30→A30, $F(2, 257) = 14$, and A60, $F(2, 257) = 79$. The quadratic component was significant in conditions A30/B30, $t(258) = 5.2$, and B30→A30, $t(258) = 5.2$, but not in condition A60, $t(258) = 0.20$.

DISCUSSION

A clear difference in the generalization gradient resulting from discriminative and non-discriminative training was again seen, and the pattern of choice probabilities and response times found was very similar to those in the last experiment. The only slight differences were in the A60 condition response times, where no quadratic component was seen, and where times seem higher for stimuli containing more than 6 B elements. In fact, the line looks closer to that found in the A30 condition of Experiment 1. This may have been due to the subjects in this experiment learning less than those in Experiment 1. Given the number who forgot the key-to-response assignment, this does not seem entirely unlikely, but, whatever the explanation, it would be wrong to become too diverted by it.

The amount of time a subject had to respond is not set in these experiments, and so response times could be subject to various uncontrolled influences (e.g. time pressure intrinsic to the subject or to experimenter effects).

The critical result is that, although there is a clear difference between discriminative and non-discriminative training, the pattern of results in the blocked and intermixed conditions is very similar. There is some evidence of a gradient difference between these conditions in the range 6–12 B elements, but it is in the opposite direction to that predicted by the incidental stimuli hypothesis. The gradient of the line in the blocked discriminative condition is shallower than in the intermixed condition in this region, whereas the non-discriminative gradient is steeper. The incidental stimulus hypothesis cannot account for the pattern of results seen in these two experiments.

GENERAL DISCUSSION

The argument put forward in the remainder of this paper is that the major patterns of results in the choice probabilities and response times of the two experiments reported can be accounted for by a simple connectionist model. The starting point for this model is the Rescorla–Wagner theory but the modification used by Gluck and Bower (1988) is not employed. This is because it is incapable of coping with situations where more than two categories need to be learned. Although this would not be a problem for these two experiments, it seemed unreasonable to constrain the model in this way. Therefore a slightly different modification was used, suggested in a footnote in the Gluck–Bower paper and previously implemented by Shanks (1990). Each category is allocated a separate node, each node having its own set of links to the stimulus components. If a category label is present, then λ for the appropriate unit is set to 1, otherwise it is set to 0.

Like the original Gluck–Bower variant, this modification predicts a difference in gradients after blocked and mixed discriminative training in the presence of incidental stimuli. If this is not completely clear, consider again the simplified situation of one perfect predictor for each category and one incidental stimulus, all of equal salience. In the first phase of the blocked condition, the perfect predictor and the incidental stimulus will become equally associated to the first category. Associations to different category representations change independently, so this is also true of the incidental stimulus and the other perfect predictor in the second phase. However, in the second phase the absence of the first category is a significant event, because the incidental stimulus is associated to it and therefore predicts that it should occur. In the continued absence of the first category, this association will extinguish. If learning is asymptotic within both phases of blocked training, then the only non-zero associative strengths are those between the first category and its perfect predictor, the second category and its perfect predictor, and the incidental stimulus and the second category—all equal to $\lambda/2$. Whether or not learning does reach asymptote, the incidental stimulus will be somewhat more associated to the second category presented than to the first, whereas in intermixed training this will not be the case. Performing the same sort of modelling demonstration as presented in Figure 4 for the Gluck–Bower variant requires that Equation 2 be modified to allow the combination of two activations. The form used in this demonstration is

$$P(A) = 1/(1 + e^{k(B-A)}) \quad 3$$

This equation is a variant of Luce's constant-ratio rule (Luce, 1959) and is formally identical to Equation 9 in Hurwitz (1994). It is also in the spirit of Shanks' derivation of predictions from his implementation (Shanks, 1991). With the same values for $\alpha \cdot \beta$ and k as used previously, this variant produces exactly the same pattern of results as shown in Figure 4. Like the Rescorla–Wagner model and the Gluck–Bower variant, this version also cannot predict the results reported here by recourse to the incidental stimuli hypothesis.

The core of the explanation offered in this paper for the difference in generalization gradients following discriminative and non-discriminative training is that the decision processes underlying categorization following discriminative training are different from those following non-discriminative training. Under discriminative training and the forced-choice "A or B" test conditions used in our experiments, it is hypothesized that the decision is basically relative ("Is this more like an A or more like a B?"). However, under non-discriminative training conditions, the decision cannot be relative in the same way. It can be done, we argue, by comparison to some minimum threshold of category membership. If this threshold is exceeded, then the subject decides that this is an example of the given category, otherwise they decide it is not. In our model this threshold is represented by a unit at the category level with an activation of a certain value not determined by stimulus information.

Any model of this form requires some way of transforming activation at category level into a prediction about how the subject will respond. Although Equation 3 fulfils this purpose, it is not used here for two reasons: First, it seems incongruous to adopt a connectionist approach and then have to "bolt on" an additional non-connectionist component to explain responding in any detail. Second, Equation 3 cannot be easily extended to explain response times. We argued earlier that some care must be taken in interpreting this part of our data because of the lack of explicit time pressure. However, the basic patterns found are reliable, and it would be informative to see whether a simple connectionist system could reproduce them.

The function of a decision rule such as that expressed in Equations 2 and 3 is to compare the activation of two or more representations and make a prediction about which one is more likely to cause its appropriate response to be made. In other words, it makes a prediction about which of the representations is more likely to win. Put in these terms, it can be seen that the same function could potentially be performed by a competitive network (Houghton, 1990; Rumelhart & Zipser, 1986). Lacouture & Marley (1991) showed that some absolute identification response time data could be predicted from the number of cycles a network of simple integrators with thresholds (similar to cascade units, McClelland, 1979) took to reach a decision. Similarly, we considered that the number of cycles in which a winner-take-all network came to a decision might be an index of response time in our categorization experiments.

Specification of the Model

The complete model is illustrated in Figure 6. Each symbol in a stimulus is represented by a single stimulus element node, which is assumed to be on if the symbol is present and off otherwise. Each category is represented by a single category node. These two sets of

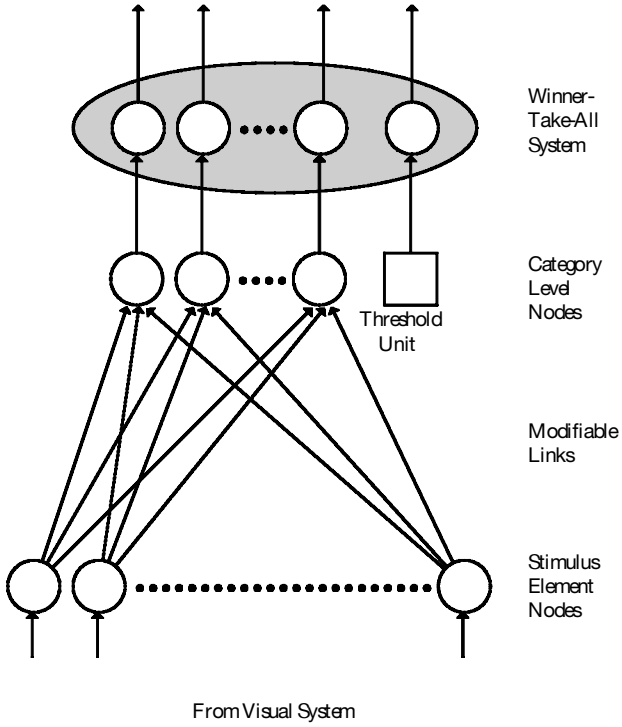

FIG. 6. The connectionist model implemented in this paper. The small dots indicate that some units have been omitted for clarity.

nodes are connected by links of modifiable strength, the modification being governed by the Rescorla–Wagner rule (Equation 1). The $\alpha\beta$ argument of this rule is replaced by a single constant α, as no assumption is made about the effect of stimulus element and category salience on the rate of learning. The asymptote of learning for links to a category node (λ) is 1 if a category label is present and 0 otherwise. The strength of all modifiable links starts at zero, and it is assumed that they change in the training phase of the experiments, but not in the test phase. The activation of a category representation is determined by the sum of the weights of links from activated stimulus units to that node. At the category level of representation there is also a threshold unit whose activation (T) is determined not by the activation of stimulus elements but by the type of decision being made.

The activations from the category level nodes are passed to a winner-take-all module via fixed links. Within this module, all nodes are self-exciting and mutually inhibitory. The activation of any unit in this system is calculated from its activation on the last cycle and the inputs coming to it. The activation of a unit on cycle c is

$$A_c = \frac{A_{c-1} + En}{1 + En + D} \qquad 4$$

if $n > 0$, and

$$A_c = \frac{A_{c-1} + En}{1 - En + D} \qquad 5$$

otherwise, where n is the total input to the unit and E and D are constants representing the rate of excitation and decay within the node. These equations are adapted from the McClelland–Rumelhart (1985) model. The values of E and D are non-critical as long as $E > D$. In the simulations performed in this paper, it is assumed that there are only two units with non-zero activation for any particular decision (the threshold unit is assumed to have zero activation on "A vs. B" decisions). Given this, n for either node can be defined as:

$$n = e + A_{c-1} - b \qquad 6$$

where e is the activation of the category node to which the winner-take-all (WTA) node is connected and b is the activation of the other WTA node. The activation of these nodes is passed on to the part of the system responsible for producing a response. This part is not explicitly modelled; instead, it is assumed that a node causes its appropriate response to be produced when its activation exceeds that of its competitor by a certain amount (d).

The system as described so far is entirely noise-free, but it seems reasonable to assume that noise would in fact be present in any neuron-like system. Noise is represented here as a random component, added at the winner-take-all stage, which has a mean of zero, a maximum level N, and a rectangular distribution. Hence e (the input from a category node) is defined as:

$$e = a + \text{rnd}(N) \cdot \text{rnd}(1, -1) \qquad 1 \geqslant e \geqslant 0 \qquad 7$$

where a is the activation of the appropriate category node, rnd(N) is a randomly determined real number from 0 to N, and rnd(1, −1) is either 1 or −1, randomly selected. Noise is assumed to be continuously varying, which is implemented here by recalculating e each cycle. Although in this instantiation noise is added only at the winner-take-all stage, this is for computational simplicity. The underlying assumption is that every stage of processing contributes some noise, and the more stages a signal has to go through, the more noise it will acquire. Hence, at the winner-take-all level, the signal from the threshold unit should be less noisy than that from the category nodes. This is because the category signal is the result of processing in many systems (including perceptual ones not represented in this model), whereas the threshold activation is an intrinsically produced signal, created at the category node level and passed straight to the winner-take-all system. Variations in noise are hypothesized to have an effect on the response-producing system. Under conditions of randomly fluctuating noise, the winner-take-all system acts like a signal amplifier—over time it amplifies the difference between signals and so reduces the effect of noise on the decision made. As overall noise increases, the signal-to-noise ratio can be maintained by increasing the activation difference threshold (d), and

we hypothesize that the response system does this. The implementation used here is to make d a function of the noise of competing units. Where there are two competing units,

$$d = s \cdot (N_1 + N_2) \quad 8$$

and s is a constant.

Before the model can be applied to the experiments reported in this paper, one further point must be made. This model assumes that the time from stimulus onset to the winner-take-all system being presented with the appropriate activations is constant. This simplification suffices, for our purposes, when the symbols presented under test are familiar. However, if the stimulus contains one or more symbols not seen in training, we assume that this involves an extra processing cost and so increases response time. For the following demonstrations, this cost is implemented by adding a constant, s_u, *to* d in situations where at least one of the elements in the presented stimulus was not seen in training. This is not intended to be a statement that the locus of the effect is necessarily in the response section of the system; it is simply an initial and imperfect instantiation of the principle that unfamiliar stimulus components will increase response time.

Simulation

The derivation of predictions from this model was performed in two stages. First, 1,000 simulated subjects were run on each of the conditions in each of the two experiments to derive the mean activation of the category units at each point in the range (12 A elements, 0 B elements) to (0 A elements, 12 B elements). Each of these sets of mean activations was then presented 1,000 times to the winner-take-all system, and the choice probability prediction derived from the number of times the Category A node won. The competition was always simulated as being between two activations—the Category A activation and either the Category B activation or the threshold activation. The index of response time was simply the mean number of cycles taken to reach a decision. The effect of stimuli with one or more elements unseen in training was simulated by estimating, in a Monte Carlo simulation, how many times such an event would occur at a given position on the range and including this many runs, within the 1,000 performed, with $d + s_u$ as the threshold for a decision. The reason for the two-stage derivation was procedural: it allowed a much easier assessment of the influence of specific variables.

The two experiments were simulated separately, but most of the model's parameters were held constant. For some this was because they were considered to be basic constants of the system (E, D, s, and N_T, the noise parameter for the threshold unit), for others because they arose somewhat from the nature of the stimuli presented (s_u and N_C, the noise parameter for category nodes). They were set as follows: $E = 0.2$, $D = 0.1$, $s = 0.2$, $N_T = 0.2$, $N_C = 1.1$, $s_u = 0.458$. Following our hypothesis that subjects in Experiment 2 learned less than did subjects in Experiment 1, we allowed α to vary across experiments. T was also allowed to vary, as it was hypothesized that the activation of the threshold unit might be dynamically altered by the system as a function of the amount of learning. For Experiment 1, $\alpha = 0.0075$, $T = 0.43$, and for Experiment 2, $\alpha = 0.0025$, $T = 0.38$. With the exception of the values for E and D, which were arbitrarily decided upon, the settings

for all parameters were chosen to allow a good approximation to the data. This was not done via any formal error minimization process, and hence the output presented is unlikely to be the best fit possible. Discovering the best-fit was not the purpose of these simulations, it was simply to demonstrate that the patterns in the data could be reproduced in some detail.

The results of the simulations are shown in Figures 7 and 8. Turning first to the choice probabilities in the simulation of Experiment 1—shown in Figure 7(a)—the overall pattern is highly similar to that seen in the data. The model correctly predicts that the gradient of the lines in the A30 and A60 conditions should be shallower than that in the A30/B30 condition in the range 0–6 B elements, and steeper in the range 6–12 B elements. To understand why this prediction is made, note that the non-discriminative lines start and end at approximately the same points as the discriminative lines, even though there is a bigger difference at these points between the activation of the A unit and the B unit than between the A unit and the threshold unit. This should lead to more certain decisions (closer to 1 or 0) in the discriminative condition at these points. The difference is compensated for by the lower noise associated with the threshold signal. Given the same start and end points, the gradient differences occur because in mid-range the difference between the threshold and the A unit activation is larger than the difference between the A and B unit activations. The model also correctly predicts that the A30 and A60 conditions are very similar, and this is due to the fact that after 30 training episodes the connection weights to the A category representation are near asymptote, and therefore the presentation of another 30 items of the same form has little effect.

Turning to the predictions of response times shown in Figure 7(b), care needs to be taken to ensure that the output is not over-interpreted. As already stated, the experimental procedures used here were not ideal for producing "clean" response times. Another, equally important, concern is that of determining the appropriate conversion from cycles to response time in seconds. Given these concerns, only the overall patterns will be discussed in any detail. First note that the model correctly predicts an inverted-U shaped trend in response times for the A30/B30 condition, approximately centred at the

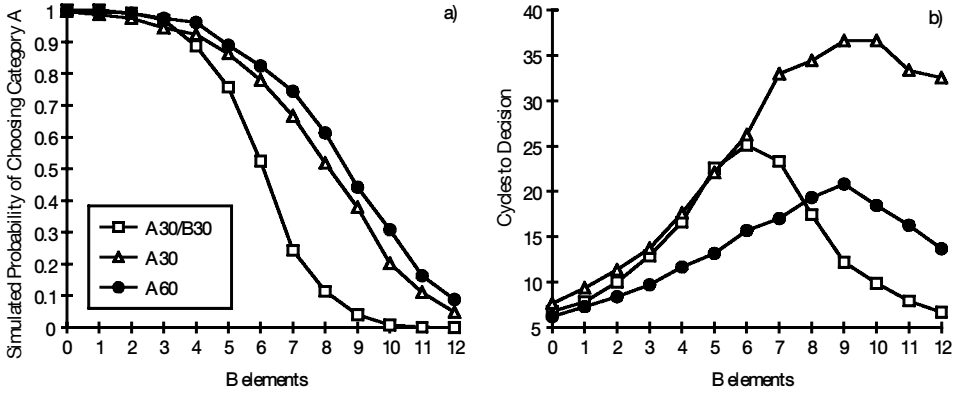

FIG. 7. Results of simulating the choice probabilities (a) and the mean number of cycles to decision (b) in the discriminative (A30/B30) and non-discriminative (A30 and A60) training conditions of Experiment 1.

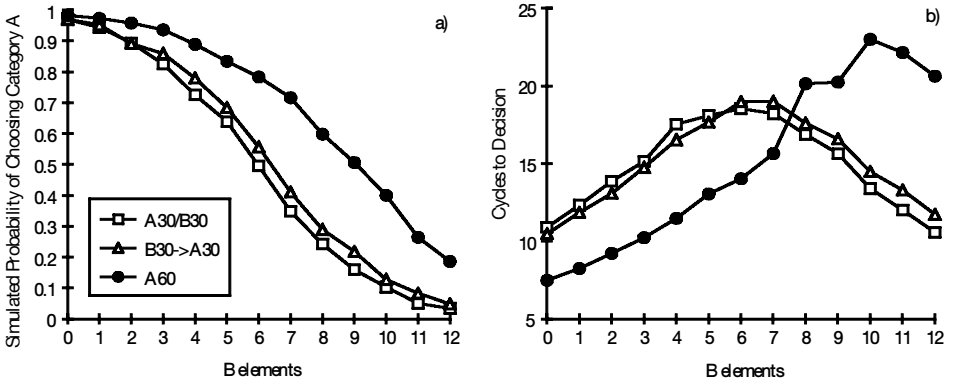

FIG. 8. Results of simulating the choice probabilities (a) and the mean number of cycles to decision (b) in the intermixed discriminative (A30/B30), blocked discriminative (B30→A30) and non-discriminative (A60) training conditions of Experiment 2.

mid-point. This trend arises because the closer two activations are, the longer, on average, it will take the system to amplify the difference above a set threshold. In the discriminative condition, the competing activations will be most similar when the amount of evidence for the items being in the two categories is equal, which, for this model, will be at 6 B elements. It would seem by this reasoning that the non-discriminative conditions should also show a similar trend. The same process is, in fact, responsible for the polynomial component of the A30 and A60 predictions, but the maximum response time is not centred at 6 B elements because this is not where the *threshold* and category activations are most similar. The predictions for the non-discriminative conditions, instead, correctly show a linearly increasing component, because the point of maximally similar activations is at about 9 B elements, and because the lower noise of the threshold unit leads to a lower response threshold, which attenuates the curvilinear component of the lines. The A30 curve is correctly shown as being generally above the A60 curve because of the effect of stimulus components unseen in training. Stimuli fulfilling this criterion are more common as the number of B elements increases, but their occurrence is so generally rare after 60 presentations that it has little effect on the A60 curve. Both the non-discriminative conditions are predicted to have some curvilinear component to their response times. Although it is true that no significant curvilinear component was found in the A30 condition, the trends in the means are not at odds with such a prediction.

The model can also correctly predict most of the patterns seen in the choice probabilities of Experiment 2—the predictions are shown in Figure 8(a). The most critical result is that a clear difference in gradient between discriminative and non-discriminative conditions can be seen in the absence of a similar difference between blocked and mixed conditions. The predictions for the non-discriminative condition differ from those for the discriminative conditions for the same reasons as they do in the predictions for Experiment 1, although the lower learning rate means that the level of noise at the threshold unit is less critical. The slight over-prediction of the probability of responding "A" in the A60 condition with stimuli containing more than 8 B elements is the by-product of the slightly lower level of T in this simulation. However, the lower value of T does allow the predic-

tion that the line will be generally above the lines in the discriminative conditions for stimuli containing 2 or more B elements.

The correct prediction that mixed and blocked discriminative training conditions are very similar is due to the absence of any incidental stimulus element representations. The small differences that remain are due to the probabilistic nature of the stimuli. Following the same argument as previously made for incidental stimuli, elements generally predictive of Category B that occur in examples of Category A in the second stage of blocked training end up somewhat associated to the Category A node. Also, their association to the Category B node decreases. The relative rarity with which any one symbol appears in a category that it is not predictive of, and the fairly low learning rate used in the modelling of both these experiments, serves to minimize the effect of this. One problem is that the differences that do remain may be in the wrong direction. The analysis of Experiment 2 suggests that, in the range 6–12 B elements, the gradient of the line in the blocked condition is significantly less negative than that in the intermixed condition. This is a pattern of results not predicted by any theory considered in this paper and would, if found to be reliable in other experiments, also pose a challenge for this model.

Finally, the model does well at predicting the pattern of response times in Experiment 2—predictions are shown in Figure 8(b). Both discriminative conditions are correctly predicted to have a marked inverted-U trend, approximately centred on the mid-point. The reason for the shape and position of the trends is the same as it is in the predictions for Experiment 1. As seen in the data, the pattern for the blocked condition is very similar to that in the mixed condition, and this is as a result of the absence of incidental stimulus component representations. The slight difference in position is due to the order effects for the associations of elements generally unpredictive of the second category to be trained (detailed in the discussion of the choice probability predictions). The non-discriminative condition shows a generally increasing component because the point of maximally similar activations is at approximately 9 B elements rather than 6 and because the curvilinear component is attenuated by the lower decision threshold caused by the lower noise of the threshold unit. The lower noise and the level of activation of the threshold are also the reason that response times for the non-discriminative condition are initially lower than for the discriminative conditions. Stimulus elements unseen in training are relatively rare in this condition and therefore have little effect.

SUMMARY AND CONCLUSIONS

When subjects acquire a novel category of the sort employed in this paper, an orderly gradient of generalization is seen around the category locus. If the category training is discriminative (involving two categories), then the gradient of generalization is sharper than if the category training is non-discriminative (involving one category). Any explanation based on incidental stimuli seems unlikely because a difference between mixed and blocked discriminative training is not seen. As far as we know, it remains an unanswered question whether similar evidence can also be found against incidental stimuli explanations of the results from rats and pigeons.

A simple connectionist model, based on a combination of the Rescorla–Wagner learning algorithm and a competitive network, has the ability to predict most of the critical patterns in choice probabilities and response times found in our two experiments. The central assumption underlying the model's success is that the decision process after discriminative training can be different from that after non-discriminative training. When only one relevant category is known about, subjects make predictions about category membership by comparison to a threshold of membership. When the decision being asked for is a forced choice between two or more categories that have been learned, this can be done by comparison of the relative levels of activation at the category nodes. Although in our experiments the stimuli used were composed of separable elements, the model we give could potentially be extended to stimuli with continuous dimensions using the stimulus coding system suggested by Shanks & Gluck (1994).

Other models, such as decision-bound theory (Ashby et al., 1994) or the Generalized Context Model (Nosofsky, 1986), may also be able to predict the pattern of results seen in these experiments. However, further development would be required before they could be applied to both the type of stimulus and experimental manipulation used here such that they could predict both the choice probabilities and the response times found. If this were done, then it may, at that point, be possible to derive divergent predictions that would allow them to be distinguished. Our model represents a first attempt to describe, in connectionist terms, the basic patterns of results found in our data, and to a large extent it is successful. The result in Experiment 2 that, in the range 6–12 B elements, the blocked discriminative gradient is shallower than the intermixed discriminative gradient is not predicted by our model. If it proves to be replicable, then some modifications may be needed.

REFERENCES

Ashby, F.G., Boynton, G., & Lee, W.W. (1994). Categorisation response time with multidimensional stimuli. *Perception & Psychophysics*, *55*, 11–27.

Buss, A.H. (1950). A study of concept formation as a function of reinforcement and generalisation. *Journal of Experimental Psychology*, *40*, 494–503.

Dennis, I., Hampton, J.A., & Lea, S.E.G. (1973). New problem in concept formation. *Nature*, *243*, 101–102.

Estes, W.K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *113*(2), 155–174.

Estes, W.K., Burke, C.J., Atkinson, R.C., & Frankmann, J.P. (1957). Probabilistic discrimination learning. *Journal of Experimental Psychology*, *54*(4), 233–239.

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247.

Grice, G.R., & Saltz, E. (1950). The generalisation of an instrumental response to stimuli varying in the size dimension. *Journal of Experimental Psychology*, *40*, 702–708.

Guttman, N., & Kalish, H.I. (1956). Generalization gradients around stimuli associated with different reinforcement schedules. *Journal of Experimental Psychology*, *51*, 79–88.

Hanson, H.M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, *58*, 321–334.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and

recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation*. London: Academic Press.

Huber, L., & Lenz, R. (1993). A test of the linear feature model of polymorphous concept discrimination with pigeons. *Quarterly Journal of Experimental Psychology*, *46B*(1), 1–18.

Hurwitz, J.B. (1994). Retrieval of exemplar and feature information in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(4), 887–903.

Jenkins, H.M., & Harrison, R.H. (1960). Effect of discrimination training on auditory generalization. *Journal of Experimental Psychology*, *59*(4), 246–253.

Jitsumori, M. (1993). Category discrimination of artificial polymorphous stimuli based on feature learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*(3), 244–254.

Lacouture, Y., & Marley, A.A.J. (1991). A connectionist model of choice and reaction time in absolute identification. *Connection Science*, *3*(4), 401–433.

Lea, S.E.G., Lohmann, A., & Ryan, C.M.E. (1993). Discrimination of five-dimensional stimuli by pigeons: Limitations of feature analysis. *Quarterly Journal of Experimental Psychology*, *46B*(1), 19–42.

Luce, R.D. (1959). *Individual choice behavior*. New York: John Wiley.

Mackintosh, N.J. (1974). *The psychology of animal learning*. London: Academic Press.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

McClelland, J.L. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287–330.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188.

Moore, J.W. (1972). Stimulus control: Studies of auditory generalisation in rabbits. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 206–230). New York: Appleton-Century-Crofts.

Newman, F.L., & Baron, M.R. (1965). Stimulus generalization along the dimension of angularity: A comparison of training procedures. *Journal of Comparative and Physiological Psychology*, *60*(1), 59–63.

Nosofsky, R.M. (1986). Attention, similarity and the identification–categorisation relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R.M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, *19*(2), 131–150.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Regehr, G., & Brooks, L.R. (1995). Category organisation in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*(2), 347–363.

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

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D.E., & Zipser, D. (1986). Feature discovery by competitive learning. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Shanks, D.R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology*, *42A*, 209–237.

Shanks, D.R. (1991). Some parallels between associative learning and object classification. In J.A. Meyer & S. Wilson (Eds.), *From animals to animats* (pp. 337–343). Cambridge, MA: MIT Press.

Shanks, D.R., & Gluck, M.A. (1994). Tests of an adaptive network model for the identification and categorization of continuous-dimension stimuli. *Connection Science*, *6*(1), 59–89.

Shepard, R. (1987). Towards a universal law of generalisation for psychological science. *Science* (September), 1317–1323.

Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13), Whole No. 517.

Original manuscript received 10 January 1996
Accepted revision received 3 January 1997