Perceptual Categorization: Connectionist Modelling and Decision Rules

F.W. Jones, A.J. Wills, I.P.L. McLaren
University of Cambridge, Cambridge, U.K.

Although it is currently popular to model human associative learning using connectionist networks, the mechanism by which their output activations are converted to probabilities of response has received relatively little attention. Several possible models of this decision process are considered here, including a simple ratio rule, a simple difference rule, their exponential versions, and a winner-take-all network. Two categorization experiments that attempt to dissociate these models are reported. Analogues of the experiments were presented to a single-layer, feed-forward, delta-rule network. Only the exponential ratio rule and the winner-take-all architecture, acting on the networks' output activations that corresponded to responses available on test, were capable of fully predicting the mean response results. In addition, unlike the exponential ratio rule, the winner-take-all model has the potential to predict latencies. Further studies will be required to determine whether latencies produced under more stringent conditions conform to the model's predictions.

As learning theorists we attempt to explain how circumstances govern behaviour. Often our explanation is couched in the form of some algorithm that, in the course of experience, mediates the relationship between stimulus input and response output. Many experiments have been performed to decide between alternative algorithms, with the aim of specifying the processes by which learning takes place. We would not wish to argue that such efforts are misplaced, but there is a danger inherent in pursuing such a strategy. If the differences observed in an experiment are not attributable to learning but, rather, stem from the operation of decision processes that operate on what has been learnt so as to convert this information into a response, then trying to account for these differences via any learning algorithm is a mistake. To avoid this type of misattribution we need to characterize the decision mechanisms sufficiently well that they can be taken into account in the interpretation of our results. That is the aim of this paper.

As an illustration of the potential pitfalls that await the unwary consider the results of Wills and McLaren (1997). In this paper the authors consider the difference between the generalization gradients found after discriminative and non-discriminative training on a

Requests for reprints should be sent to F.W. Jones, The Psychological Laboratory, University of Cambridge, Downing Street, Cambridge, CB2 3EB, U.K. E-mail: fwj1000@cus.cam.ac.uk

This research was funded by a BBSRC grant to I.P.L. McLaren. The authors would like to thank Howard Portess, who assisted in running the first experiment, and Mike Aitken, Martin Redington, and two anonymous reviewers for their many helpful comments.

categorization task with human subjects. The result is that the post-training gradients are significantly steeper in the discriminative case, an effect that parallels that found in the animal learning literature. The explanation offered for this effect in pigeons is that discriminative training allows the features relevant to the discrimination to overshadow (incidental) stimuli common to S+ and S−, so lessening generalization between them and giving a steeper gradient, a type of relative validity effect (Wagner, Logan, Haberlandt, & Price, 1968). Wills and McLaren were able to show that this was not the explanation for the effect in their experiments, however, and that it could be explained as a consequence of the interplay between decision mechanism and task requirements. Put simply, in the discriminative case some comparison is made of alternatives, whereas in the non-discriminative case comparison is made with respect to some criterion. Wills & McLaren were able to show that this followed from a connectionist model of the decision process that did not rely on the overshadowing mechanism.

There is no reason to suppose that there could not be a decision mechanism of this kind in the pigeon, in which case attribution of the discriminative/ non-discriminative gradient difference to an overshadowing mechanism in all circumstances may be unwarranted. We simply don't know. The reason for this is that the necessary experiments deconfounding the effects of decision processes and overshadowing have not been done. This is symptomatic of a general tendency to neglect the whole issue of response selection by comparison with the study of learning itself. We model learning but stop short of specifying how it results in action.

In general, if we follow the connectionist approach to modelling, then a decision process that converts output activations to probabilities of response is required. This can take the form of a function known as a decision rule, which in many cases is implicit in the modelling rather than an explicit component of the model. Here we consider four candidate functions that can be used to model the decision process.

The simple ratio rule is formally stated as:

$$P(a) = A_a / \Sigma A_n \qquad 1$$

where $P(a)$ is the probability of responding "a", A_x is the activation of the appropriate category unit(s), and n is the number of categories being considered. This rule is preferred by Shanks (e.g. Shanks, 1991) and Estes (1950). It corresponds to the idea that the probability of selecting one of a number of possible choices depends on its "weight" compared to those alternatives. As such, it has something in common with the notion of matching response probabilities to associative strengths.

A variant of this is the exponential ratio rule, which is expressed thus:

$$P(a) = e^{kA_a} / \Sigma e^{kA_n} \qquad 2$$

where k is a free parameter. The inclusion of the exponential transform allows the function to be non-linear and constrains the terms to positive values. Authors who employ this rule include McClelland and Rumelhart (1981, 1985), McClelland and Elman (1986), and Gluck and Bower (1988). The popularity of ratio rules probably stems from Luce's Choice Axiom (Luce, 1963) and the evidence that is consistent with it (e.g. Bradley, 1954;

Hopkins, 1954). This exponential variant is more common in modelling human decision processes.

An alternative class of mechanism considered here is exemplified by the simple difference rule:

$$R(a) = [k(A_a - A_b) + 1]/2 \qquad 3$$

where just two categories, a and b, are being considered. McLaren, Bennett, Guttman-Nahir, Kim, & Mackintosh (1995) discuss a rule of this type. Modification of this gives the exponential difference rule:

$$R(a) = (e^{kA_a} - e^{kA_b} + e^{k})/e^{2k} \qquad 4$$

In both cases the numerical constants serve to fit the functions to an appropriate range. The application of difference rules to situations involving more than two categories is ambiguous. However, one posibility is to compare the two categories with the most active units. Difference rules have been thought to be plausible candidates because under certain circumstances they might naturally emerge from a competitive response stage. In such a process the probability of a particular response would depend upon the advantage (i.e. difference) in activation of its unit over the other units. This type of rule is implicit in a great deal of work modelling learning. Whenever an author points out that the associative strength for one stimulus exceeds that for another by more than for a comparable set of stimuli in the experiment, it is implied that it is the difference in associative strengths rather than their relative weightings that controls performance.

The possibility of modelling the decision process using a connectionist instantiation of a winner-take-all response stage, similar to that employed by Wills and McLaren (1997), is also examined in this paper. A model of this type has the advantages of general applicability and being able to produce predictions concerning both latency and probability of response. However, a detailed description of the network simulations will be left until later.

A preliminary analysis of the behaviour of the decision rules indicated that for a case involving just two category units, a halving of both their activations did not alter the probabilities of response predicted by the simple ratio rule, but did affect those produced by the other rules. These differential predictions provided the basis for the first experiment.

EXPERIMENT 1

The stimuli employed comprised a 4 × 3 array of spatially separate symbols whose position conveyed no information; the symbols were of the same form as those used by Wills and McLaren (1997). Their elemental nature allowed them to be easily coded in a connectionist network. Furthermore, it was expected that the large size of the arrays would make it difficult for subjects initially to adopt a cognitive, rule-based strategy. The training procedure employed was also similar to that used by Wills and McLaren. Specifically, instead of learning through feedback to their responses, subjects merely had to observe while the appropriate category information was shown in conjunction with each exemplar.

The experiment comprised two conditions: $A_{30}B_{30}$, in which the training exemplars, which were distortions of two non-overlapping prototypes, A and B, were presented with their correct category labels; and $A_{15}B_{15}C_{30}$, in which a random half of the training exemplars from each category were incorrectly assigned to an alternative category, C. The latter condition can be considered as implementing a form of partial reinforcement during training. For both conditions training was followed by a binary categorization task, AB, in which a series of test exemplars spanning the range between the two prototypes had to be classified as either A or B. In addition, the $A_{30}B_{30}$ condition included a second test manipulation, ABX, which employed a series of test exemplars that were constructed identically to the AB set except that half of their elements were replaced with ones that had not previously been seen. This last test condition can be thought of as a form of generalization decrement manipulation.

Thus, there were three experimental groups: $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX, which were compared within subject because of their identical training phases, and $A_{15}B_{15}C_{30}$/ AB, which employed different subjects. The latter two were designed such that when modelled using a single-layer feed-forward architecture, their output activatons on test were half the magnitude of those produced for corresponding positions between the prototypes by the $A_{30}B_{30}$/ AB group. This modelling is considered in detail after the presentation of the results. Both the $A_{30}B_{30}$/ ABX and the $A_{15}B_{15}C_{30}$/ AB groups were included in the hope of providing convergent evidence on the effects of degrading the information needed to categorize the stimuli.

Method

Subjects and Apparatus

The subjects were 32 adults, mostly Cambridge undergraduate students, whose ages ranged between 21 and 55. Half were tested on the $A_{30}B_{30}$/ AB + ABX condition, and the rest undertook the $A_{15}B_{15}C_{30}$/ AB condition.

The experiment was run from three Acorn RiscPC 600 computers with colour monitors, each in a quiet room. The instructions and stimuli were presented on screen, and the subjects responded by pressing keys on the keyboard. Responses and latencies were logged in a data file.

Stimuli

Each stimulus consisted of 12 different symbols randomly arranged on an invisible 4 × 3 grid. In total 36 symbols were employed. These are shown together with an example stimulus in Figure 1.

The stimuli for both the $A_{30}B_{30}$ and the $A_{15}B_{15}C_{30}$ training phases were generated in an identical manner. For every subject, 30 of the 36 symbols were arbitrarily selected. These were randomly divided into two groups of 12, one for each prototype A and B, and a set of 6 not used in training but required for the ABX test phase. Exemplars were constructed by assigning a 10% chance of exchange to each symbol of the relevant prototype. Exchange consisted of the replacement of the symbol by one randomly selected from the other prototype, with the caveat that all the replacement symbols had to be different. The symbols from each resulting exemplar were allocated arbitrary locations on an invisible 4 × 3 grid. In this manner 30 training exemplars were produced for each category.

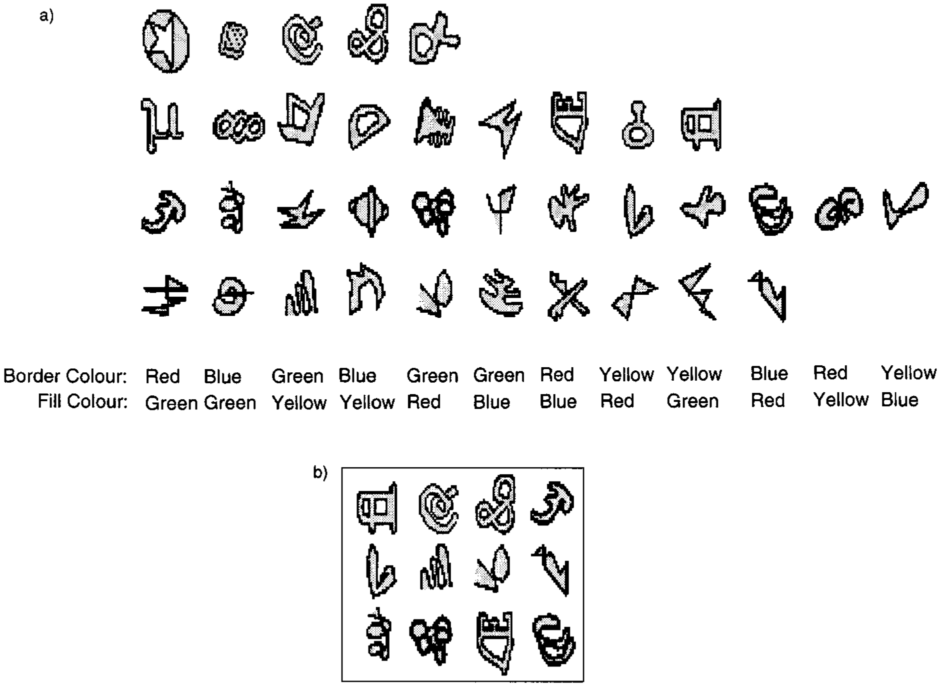
a)
Border Colour: Red Blue Green Blue Green Green Red Yellow Yellow Blue Red Yellow
Fill Colour: Green Green Yellow Yellow Red Blue Blue Red Green Red Yellow Blue
b)

FIG. 1. (a) The 36 symbols employed in the experiment, arranged by colour (b) An example stimulus.

For every subject a set of AB test exemplars was constructed to cover systematically the range between prototype A (12 A symbols, no B symbols) and prototype B (no A symbols, 12 B symbols). Henceforth these exemplars will be described in terms of their "proportion of B symbols", which is the ratio of the number of B symbols present to the total number of A and B symbols. Five exemplars were created for each of 7 different values of this ratio, namely 0/12 (Prototype A), 2/12, 4/12, 6/12, 8/12, 10/12, and 12/12 (Prototype B). Each was constructed by randomly selecting the appropriate numbers of A and B symbols, such that no two symbols were the same, and arbitrarily arranging the result on an invisible 4 × 3 grid. Stimuli for intermediate values of the ratio—for example, 1/12—were not created as it would not be possible to compare them to exemplars from the ABX test phase.

ABX test stimuli differed from AB exemplars in that half their constituent symbols were ones previously unseen by the subject. They therefore ranged from 6 A symbols, no B symbols, and 6 novel symbols, to no A symbols, 6 B symbols, and 6 novel symbols. For purposes of comparison, 5 exemplars were created for each of the 7 "proportions of B symbols" used to generate the AB stimuli, namely 0/6, 1/6, 2/6, 3/6, 4/6, and 6/6. In all other respects the manner of construction was identical to that described above. The composition of all the stimuli from both test phases is summarized in Table 1.

Design

Condition $A_{30}B_{30}$/AB+ABX: In the training phase subjects were presented with the 60 training stimuli successively, in an arbitrary order. Each stimulus was labelled either A or B, depending upon its category. No response was required from the subjects. For the test phase the 70 AB and ABX test stimuli were randomly intermixed and then displayed consecutively without labels. The subjects

TABLE 1
The Proportions of A, B, and Novel Symbols Used to Construct the Various Test Exemplars, for Both the AB and ABX Test Phases

	AB Test Exemplars			*ABX Text Exemplars*		
Proportion of B symbols	*Number of A symbols*	*Number of B symbols*	*Number of novel symbols*	*Number of A symbols*	*Number of B symbols*	*Number of novel symbols*
0/ 12	12	0	0	6	0	6
2/ 12	10	2	0	5	1	6
4/ 12	8	4	0	4	2	6
6/ 12	6	6	0	3	3	6
8/ 12	4	8	0	2	4	6
10/ 12	2	10	0	1	5	6
12/ 12	0	12	0	0	6	6

were directed to categorize each stimulus as either an A or a B and encouraged to respond as quickly as convenient while being as accurate as possible.

Condition $A_{15}B_{15}C_{30}$/ AB: In the training phase a randomly selected half of the training stimuli from each category were assigned the label C, and the rest were correctly designated. To reduce the likelihood of subjects learning that C stimuli were in fact exemplars from the other two categories, they were never displayed either immediately before or after stimuli from their true category. The order of presentation was an arbitrary combination of the following sequences:

4 × A C_b A B C_a B C_a C_b
7 × A C_b C_a B

where A and B represent correctly labelled training stimuli, and C_a and C_b denote C stimuli and their original category. The test phase was identical to that described above, except that it only involved the AB test stimuli.

Procedure

Each subject was requested to read the general instructions presented on the computer screen. These comprised a description of the whole experiment, some details concerning the first phase, and an example of a training stimulus with accompanying label. The experimenter verified that the subject understood the instructions and then left the room for the remainder of the experiment. The subject initiated the training phase by pressing any key on the keyboard. Each of the stimuli, which were approximately 4.5 cm × 3.5 cm, appeared in the centre of the screen. The appropriate letter label—an A, B, or C—which was of equivalent size to the stimulus appeared to its right. The subjects were seated approximately 1 m away from the computer screen, which was approximately at eye level. Each stimulus–label pair was presented for 5 sec. This was followed by a plain grey mask, which covered the location of both the stimulus and the label, that appeared for 2 sec. The next pair was then shown.

After all the training stimuli had been presented, details concerning the test phase were displayed. Each subject was instructed to classify every test stimulus as either an A or a B as soon as possible but with the emphasis on accuracy, and informed of the appropriate key to press for each case. The subject initiated the test phase by pressing any key on the keyboard. Each stimulus was presented until the subject had responded, whereupon the next exemplar appeared. This time the stimuli were not labelled. A reminder of the response key assignments was always present on the screen. Pressing any key other than the two that were designated caused the computer to issue a beep and wait for an appropriate response. Upon completion of the experiment the computer automatically stored the results, both responses and latencies, in a data file. Each subject was thanked for participating.

Results

Two measures of performance on test trials—response and latency—were taken.

Responses

A plot of the subjects' mean response against proportion of B symbols for all three groups is shown in Figure 2a. ANOVAs were conducted on the means from each subject. Given that groups $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX were combined, whereas the $A_{15}B_{15}C_{30}$/ AB group employed different subjects, the groups were analysed in pairs using three separate ANOVAs. Group was a within-subject variable for the comparison of groups $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX, otherwise it was a between-subject variable.

Groups $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX Compared: Proportion of B symbols had a significant effect on response, $F(6, 90) = 72.08$, $p < 0.05$, whereas group did not, $F(1, 15) = 2.69$. However, there was a significant interaction between group and proportion of B symbols, $F(6, 90) = 4.99$, $p < 0.05$.

Groups $A_{30}B_{30}$/ AB and $A_{15}B_{15}C_{30}$/ AB Compared: Proportion of B symbols exerted a significant effect, $F(6, 180) = 57.73$, $p < 0.05$, group did not $F(1, 30) < 1$, but the interaction was significant, $F(6, 180) = 6.40$, $p < 0.05$.

Groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB Compared: Again the effect of proportion of B symbols was significant, $F(6, 180) = 24.63$, $p < 0.05$, whereas that of group was not, $F(1, 30) = 2.74$. However, there was no significant interaction between the two, $F(6, 180) = 1.14$.

Linear regression was employed to approximate the results from each group with a straight line. In all cases the model fit was significant, Fs$(1, 110) > 56.29$, $p < 0.05$. In order to determine whether there were significant differences in gradients between groups, separate regression lines were calculated from each subject's data. The gradients for groups $A_{30}B_{30}$/ AB (mean = −0.0853) and $A_{30}B_{30}$/ ABX (mean = −0.0500) differed significantly, $t(30) = 4.86$, $p < 0.05$, as did those for groups $A_{30}B_{30}$/ AB and $A_{15}B_{15}C_{30}$/ AB (mean = −0.0440), $t(30) = 3.95$, $p < 0.05$. However, there was not a significant difference when groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB were compared, $t(30) = 0.50$.

Latencies

Latency should be regarded as a secondary performance measure to response, as the subjects were not required to respond within a limited time and so were under little time pressure. The mean latencies for subjects in each group are shown against proportion of B symbols in Figure 2b. ANOVAs were performed on the means from each subject.

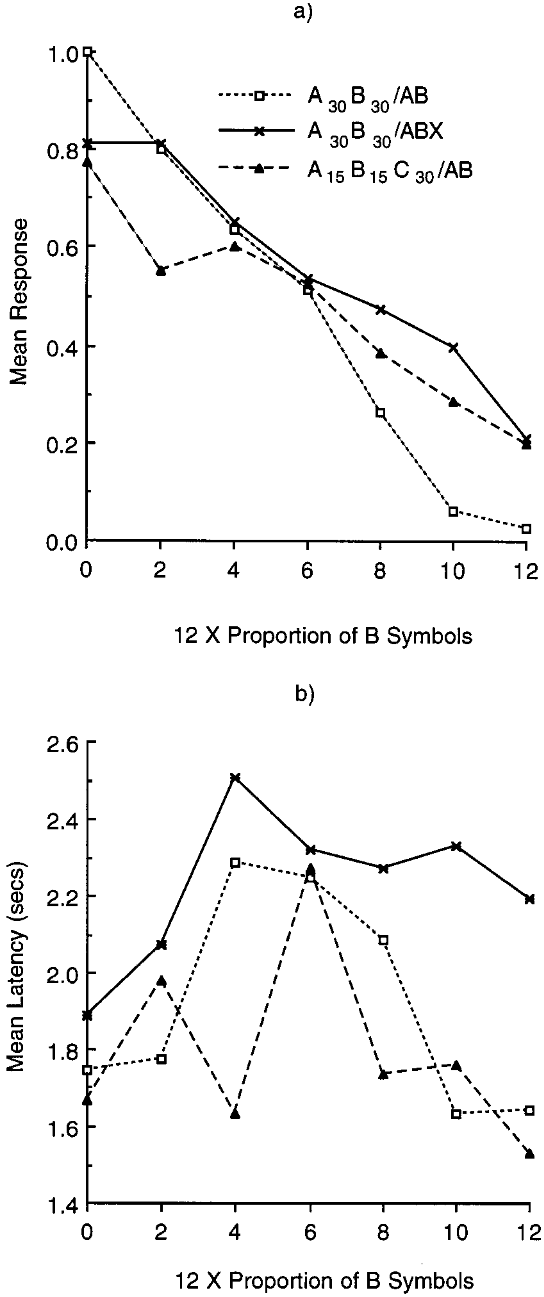

FIG. 2. (a) Mean response data for groups $A_{30}B_{30}$/ AB, $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB. Mean response is equivalent to the probability of responding "A". (b) Mean latency, in seconds, for groups $A_{30}B_{30}$/ AB, $A_{30}B_{30}$/ ABX, and $A_{15}B_{15}C_{30}$/ AB.

Groups $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX Compared: Subjects took significantly longer to respond in group $A_{30}B_{30}$/ ABX than in group $A_{30}B_{30}$/ AB, $F(1, 15) = 17.40$, $p < 0.05$, and proportion of B symbols had a significant effect on mean latency, $F(6, 90) = 3.53$, $p < 0.05$. However, there was no significant interaction between group and proportion of B symbols, $F(6, 90) = 1.04$.

Groups $A_{30}B_{30}$/ AB and $A_{15}B_{15}C_{30}$/ AB Compared: Proportion of B symbols exerted a significant effect, $F(6, 180) = 4.44$, $p < 0.05$, but group did not, $F(1, 30) < 1$. Nor was the interaction significant, $F(6, 180) = 1.95$.

Groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB Compared: Again the effect of proportion of B symbols was significant, $F(6, 180) = 2.46$, $p < 0.05$, and that of group was not, $F(1, 30) = 2.48$. The interaction between the two just failed to reach significance $F(6, 180) = 2.21$.[1]

Discussion

This experiment confirms Wills and McLaren's conclusions; first, that people do not necessarily find it difficult to learn to distinguish categories with polymorphous-like structure, and second, that an orderly generalization gradient is produced by discriminative training. However, for present purposes the interesting finding is that group $A_{30}B_{30}$/ AB has produced a steeper generalization gradient than groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB, whereas the gradients generated by the latter are effectively the same. This suggests that what counts are the activations of (or net associative strengths to) the output units corresponding to the available responses. From a theoretical perspective, it is in this sense that the "partial reinforcement" and "generalization decrement" conditions are equivalent.

Modelling

In order to establish which decision rules were capable of predicting this result, an analogue of the experiment was presented to a single-layer, feed-forward, delta-rule network in combination with the various decision rules. Given that the categories were linearly separable, it was not felt necessary to employ a more complex architecture using an algorithm such as backpropagation.

The network comprised 30 feature (input) units and three name (category) units that each received connections from all the feature units. The activation of a name unit was calculated by taking the weighted sum of the activations of all the units connected to it. Formally, for name unit n connected to F feature units:

$$a_n = \sum_F a_f w_{nf} \qquad 5$$

[1] Although $F(6, 180) = 2.21$ is actually a significant value, when the degrees of freedom were adjusted in accordance with the Greenhouse and Geisser (1959) correction for non-sphericity, giving $F(4.86, 145.8) = 2.21$, it no longer reached significance. This correction did not affect any of the other analyses reported.

where w_{nf} is the weight of the connection from feature unit f to name unit n, and a_x is the activation of unit x. Initially the weights were set to zero, and learning followed the delta rule (McClelland & Rumelhart, 1985). Formally:

$$\delta w_{nf} = S\Delta_n a_f \qquad \text{given that} \qquad \Delta_n = e_n - a_n \tag{6}$$

where e_n is the external input to name unit n, and S is a constant that determines the rate of learning. S was set to 0.005 for all the simulations.

Each feature unit corresponded to a particular symbol. Thus the experimental stimuli were represented by setting to one the activation of those feature units whose symbols were present and assigning a value of zero to the rest. The three name units corresponded to the three categories A, B, and C. For each training exemplar the external input to the name unit representing its category was set to one, and the rest took a value of zero. The weights were updated once by the delta rule after the presentation of each training stimulus. During test the weights were frozen and the activations of the name units produced by each exemplar recorded.

In all other respects both the generation and order of presentation of the stumuli were identical to that in the experiment. Thus the $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX groups were combined and run together on the same network, whereas group $A_{15}B_{15}C_{30}$/ AB employed a new network.

The mean activations produced on test for each proportion of B symbols and every group were calculated from the results of 1,000 simulations. The four decision rules were applied to the A and B name unit values from this data. The exclusion of the C unit values is of no relevance to groups $A_{30}B_{30}$/ AB and $A_{30}B_{30}$/ ABX, because there were no C exemplars in these groups, resulting in the C unit activations always being zero.

The probabilities of response produced by the various rules are shown in Figures 3 and 4. They are taken to be indices of the mean response measures used in the experiment. The values of k employed were chosen to be consistent with the modelling of the second experiment, which will be discussed later. It can be seen that all the rules, except the simple ratio rule, successfully predict that the generalization gradient produced by the $A_{30}B_{30}$/ AB group should be steeper than that for group $A_{30}B_{30}$/ ABX.

For the $A_{15}B_{15}C_{30}$/ AB group, which had a training phase that included C exemplars, the fact that C unit values were omitted did affect the predictions of the various rules. Their omission was based on the assumption that as on test subjects had no opportunity to respond “C”, their performance would be adequately described by a decision process based only on A and B unit activations. The modelling predicted that if this assumption was correct, then the mean responses for groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB should not differ significantly, as their A and B name unit activations were effectively equivalent.[2] The experimental results demonstrated this to be the case, thus justifying the assumption. From Figures 3 and 4 it is clear that the simple ratio rule is the only function inconsistent

[2] In fact, there was a small difference between the A and B name unit activations for groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB, which was not entirely due to the randomized aspects of the simulations. Instead, it was a result of the constraints placed upon the order of presentation of the training exemplars in group $A_{15}B_{15}C_{30}$/ AB. However, given the very small magnitude of this effect, it is not surprising that it does not reach significance in the empirical results, and it does not seem to be a matter for concern.

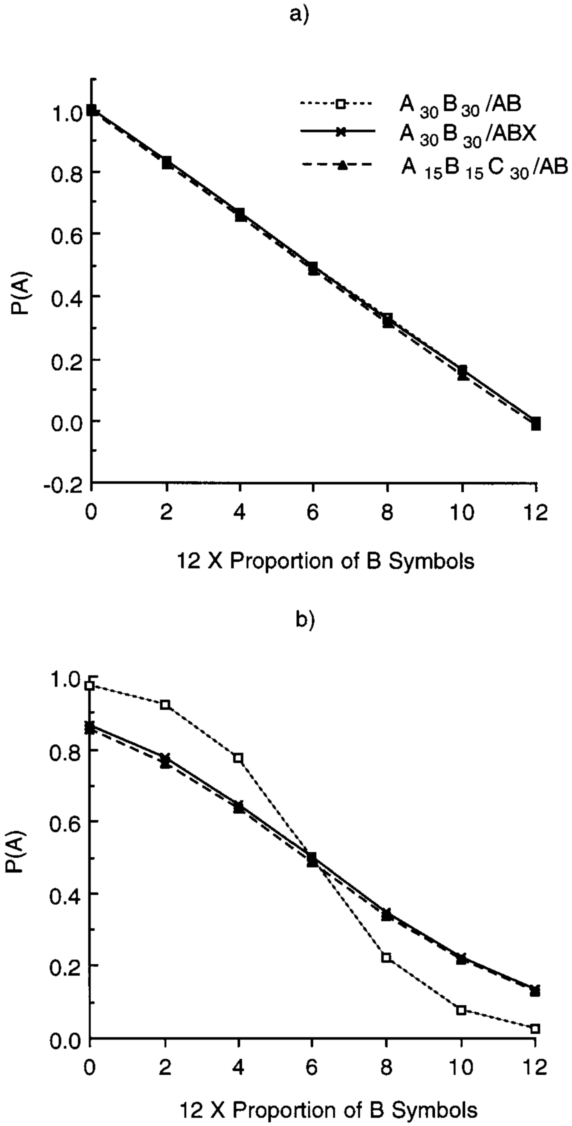

FIG. 3. Probabilities of responding "A" predicted by (a) the simple ratio rule, and (b) the exponential ratio rule for $k = 4.4$.

with the finding that the generalization gradient for the $A_{15}B_{15}C_{30}$/ AB group was shallower than that produced by group $A_{30}B_{30}$/ AB.

In summary, the evidence converges on the conclusion that the simple ratio rule does not provide an adequate model of human performance, at least with regard to the type of

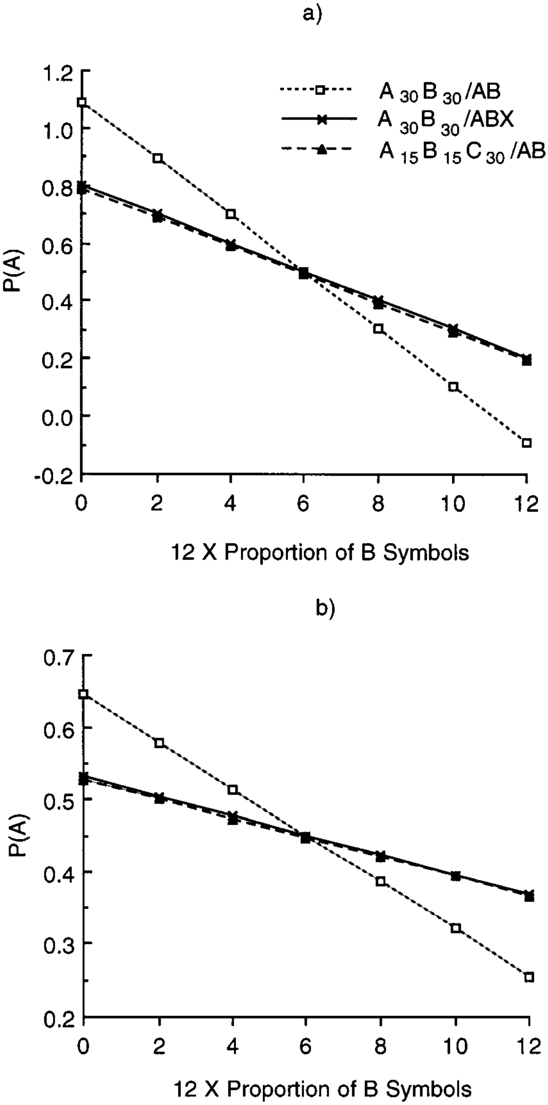

FIG. 4. Probabilities of responding "A" predicted by (a) the simple difference rule for $k = 1.4$, and (b) the exponential difference rule for $k = 0.8$.

task and network considered here. The fact that the mean response results for groups $A_{30}B_{30}$/ ABX and $A_{15}B_{15}C_{30}$/ AB were very similar suggests that, at least in this case, decision rules should only include the activations of output units that correspond to responses available on test. A discussion of the latency data will be left until the winner-take-all architecture is considered, as the decision rules used here only generate predictions concerning mean response.

In order to dissociate the three remaining candidate decision rules, a second experiment was performed. This experiment invokes another method of manipulating output activation—the amount of training given.

EXPERIMENT 2

The form of the task and stimuli employed were the same as in Experiment 1. Experiment 2 comprised two groups: $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB*. Group $A_{30}B_{30}$/ AB* was identical to group $A_{30}B_{30}$/ AB in all except two respects: first, during the test phase subjects were presented with 10 exemplars for each proportion of B symbols rather than 5, with the aim of reducing the variability in the data; second, the test exemplars covered all 13 possible proportions of B symbols instead of just 7. Thus a more detailed generalization gradient was produced so increasing the chance of discriminating between the rules.

Group $A_{10}B_{10}$/ AB* differed from group $A_{30}B_{30}$/ AB* in that subjects only received 10 training exemplars per category in the former compared to 30 per category in the latter. Given this difference in training, the two groups were a between-subjects manipulation. These groups were chosen so that when modelled using the single-layer feed-forward network, the former produced output activations on test approximately half the magnitude of those generated for corresponding proportions of B symbols by the latter.

Method

Subjects and Apparatus

The subjects were 32 adults, mostly Cambridge graduate students, who were paid for their participation. Their ages ranged between 17 and 53. None had been included in Experiment 1. Half were run on group $A_{30}B_{30}$/ AB* and the remainder on group $A_{10}B_{10}$/ AB*. The apparatus was the same as that employed in Experiment 1, apart from the fact that only one computer was used.

Stimuli

The stimuli for both groups were generated in an identical fashion to those for group $A_{30}B_{30}$/ AB, except that test exemplars were constructed for all 13 possible proportions of B symbols.

Design

Group $A_{30}B_{30}$/ AB:* In the training phase 60 labelled exemplars, half from each category, were presented to the subjects successively, in an arbitrary order. During the test phase 130 exemplars, 10 for each proportion of B symbols, were randomly intermixed and displayed consecutively. As in Experiment 1, the subjects were required to categorize each test exemplar as either an A or B and were requested to respond as soon as possible consistent with making as few errors as possible.

Group $A_{10}B_{10}$/ AB^*: The training phase differed from that in group $A_{30}B_{30}$/ AB* by the fact that only 10 exemplars from each category were presented. Testing was identical to that for group $A_{30}B_{30}$/ AB*.

The procedure was the same as that employed in Experiment 1.

Results

Responses

Figure 5a shows subjects' mean responses plotted against proportion of B symbols for both groups. An ANOVA was conducted on the means from each subject, with group as a between-subject variable. Proportion of B symbols was found to have a significant effect on response, $F(12, 360) = 87.53$, $p < 0.05$, but group did not, $F(1, 30) < 1$. The interaction between group and proportion of B symbols was significant, $F(12, 360) = 2.55$, $p < 0.05$.

Linear regression demonstrated that both groups could be meaningfully modelled using straight lines, $F(1, 206) > 284.49$, $p < 0.05$. As for Experiment 1, separate regression lines were calculated from each subject's data. It was found that group $A_{30}B_{30}$/ AB* (mean $= -0.0967$) had a significantly steeper gradient than did group $A_{10}B_{10}$/ AB* (mean $= -0.0734$), $t(30) = 3.35$, $p < 0.05$.

Latencies

Again it should be appreciated that because the subjects were not under severe time pressure, latency can only be treated as a secondary performance measure. The mean latencies for subjects in both groups are shown against proportion of B symbols in Figure 5b. An ANOVA was performed on the means from each subject. This established that group did not have a significant effect on mean latency, $F(1, 30) < 1$, but proportion of B symbols did, $F(12, 360) = 7.97$, $p < 0.05$. There was no significant interaction between the two, $F(12, 360) < 1$.

Discussion

From the results it is clear that group $A_{30}B_{30}$/ AB* produced a steeper generalization gradient than did group $A_{10}B_{10}$/ AB*—a result that is very much in line with those in Experiment 1. The experiment was simulated in order to determine which of the remaining decision rules were capable of accounting for this finding.

Modelling

The groups were modelled in a manner identical to that used for Experiment 1, except that the C name unit was omitted as the experiment had not included C exemplars. The probabilities of response generated by the various rules are presented in Figure 6. All three rules successfully predicted that the $A_{30}B_{30}$/ AB* group's generalization gradient should be steeper than that for the $A_{10}B_{10}$/ AB* group.

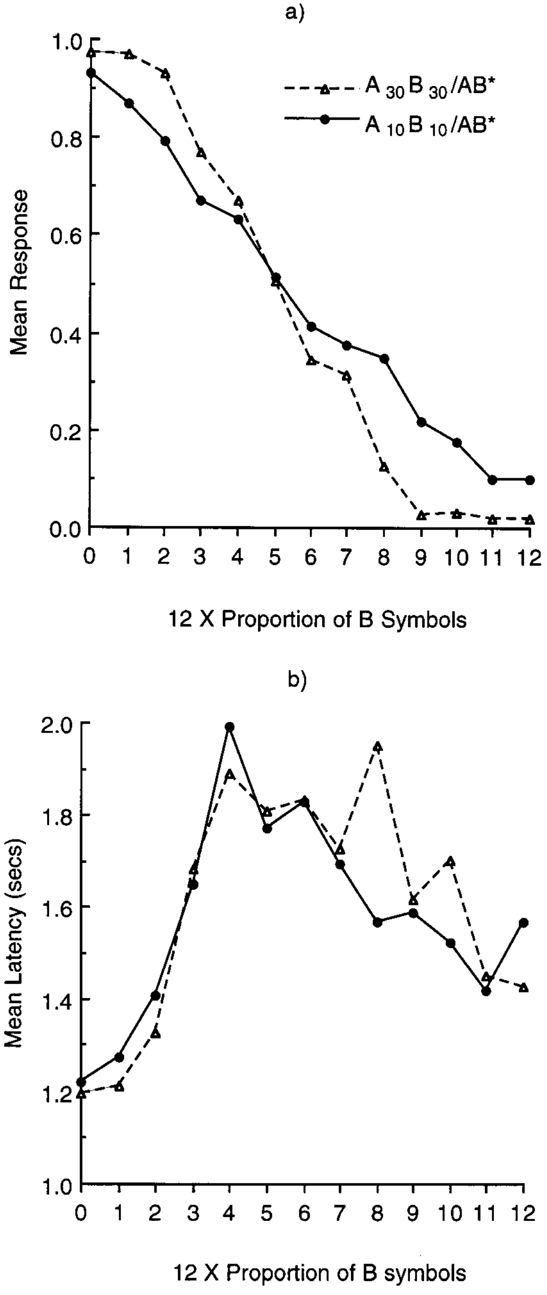

FIG. 5. (a) Mean response data for groups $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB*. (b) Mean latency, in seconds, for groups $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB*.

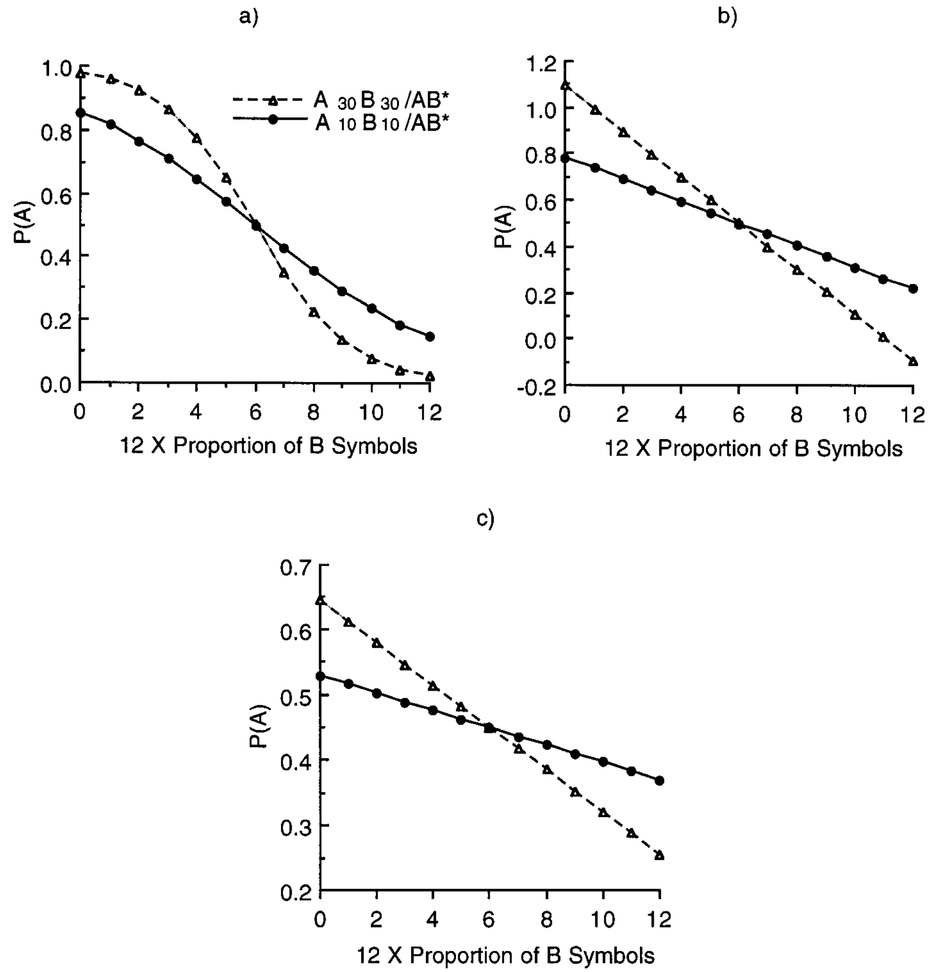

FIG. 6. Probabilities of responding “A” predicted by (a) the exponential ratio rule for *k* = 4.4, (b) the simple difference rule for *k* = 1.4, and (c) the exponential difference rule for *k* = 0.8.

Hence, in an attempt to dissociate between them, a more detailed comparison of their predictions and the $A_{30}B_{30}$/ AB* group’s mean response data was performed. Group $A_{30}B_{30}$/ AB* was chosen because, first, it had a better-resolved generalization gradient than those produced by groups $A_{30}B_{30}$/ AB, $A_{30}B_{30}$/ ABX, and $A_{15}B_{15}C_{30}$/ AB, and, second, the sigmoidal shape of its curve was more pronounced than that for group $A_{10}B_{10}$/ AB*. Numerical minimization was employed to find for each rule the value of *k* that produced the smallest mean squared error between its predictions and the means from each subject. The values for *k* were 4.4, 1.4, and 0.8, for the exponential ratio rule, simple difference rule, and exponential difference rule, respectively. For purposes of consistency they have been used throughout this paper.

From Figure 7a it can be seen the predictions of the exponential ratio rule always lie within the envelope described by the 5% error bars, corrected for the number of comparisons, of the mean response data. However, Figures 7b and 8a demonstrate that this is not the case for either the exponential difference rule or the simple difference rule, which have, respectively, nine and two points outside the error bars. As a correction has been made for the number of comparisons, the predictions of a given rule can be said to be significantly different from the data even if they fall outside the error bar for only one point. Therefore the probabilities of response produced by the exponential ratio rule are the only ones not significantly different from the empirical results for group $A_{30}B_{30}$/ AB*.

Clearly the exponential difference rule is not a feasible candidate to model the decision process. However, it might still be argued that it would not be justified to reject the simple difference rule when it lies marginally outside the error bars for only two of the thirteen points. Therefore the rule was modified to determine whether the inclusion of a second free parameter, *g*, would allow it to account for the results. Formally:

$$P(a) = [k(A_a - A_b) + 1]/2 + g \qquad 7$$

g can be considered to be a factor to compensate for subjects' preference for one particular response key. However, as Figure 8b shows, even with the addition of an extra free parameter, after best fitting one point still lies outside the error bars of the mean response data.

There are two additional problems with the simple difference rule. First, since it always produces a straight line regardless of the magnitude of *k*, it totally fails to capture the sigmoidal shape of the data. Second, it predicts impossible probabilities of response—namely, ones above one and below zero. Moreover, if the free parameter(s) were adjusted to prevent this, then its fit to the group $A_{30}B_{30}$/ AB*, data would become even worse. Given all these difficulties, it can, at the very least, be concluded that the exponential ratio rule provides a substantially better model of the decision process than does the simple difference rule.

GENERAL DISCUSSION

In summary, the exponential ratio rule, applied to those activations that correspond to responses available on test, generates the most satisfactory description of human performance for the type of task and network considered here. However, it does not produce predictions concerning latencies, nor does it offer a neuron-like mechanism for the decision process. An alternative that satisfies both of these criticisms is to model the decision stage using a noisy winner-take-all network.

Modelling

Figure 9 illustrates the architecture employed. It consisted of two units, one for each possible response. Both units had a self-excitatory connection, an excitatory link from the appropriate name unit, and an inhibitory connection from the other response unit. The weights of these links were fixed. Upon the presentation of the name unit activations, the

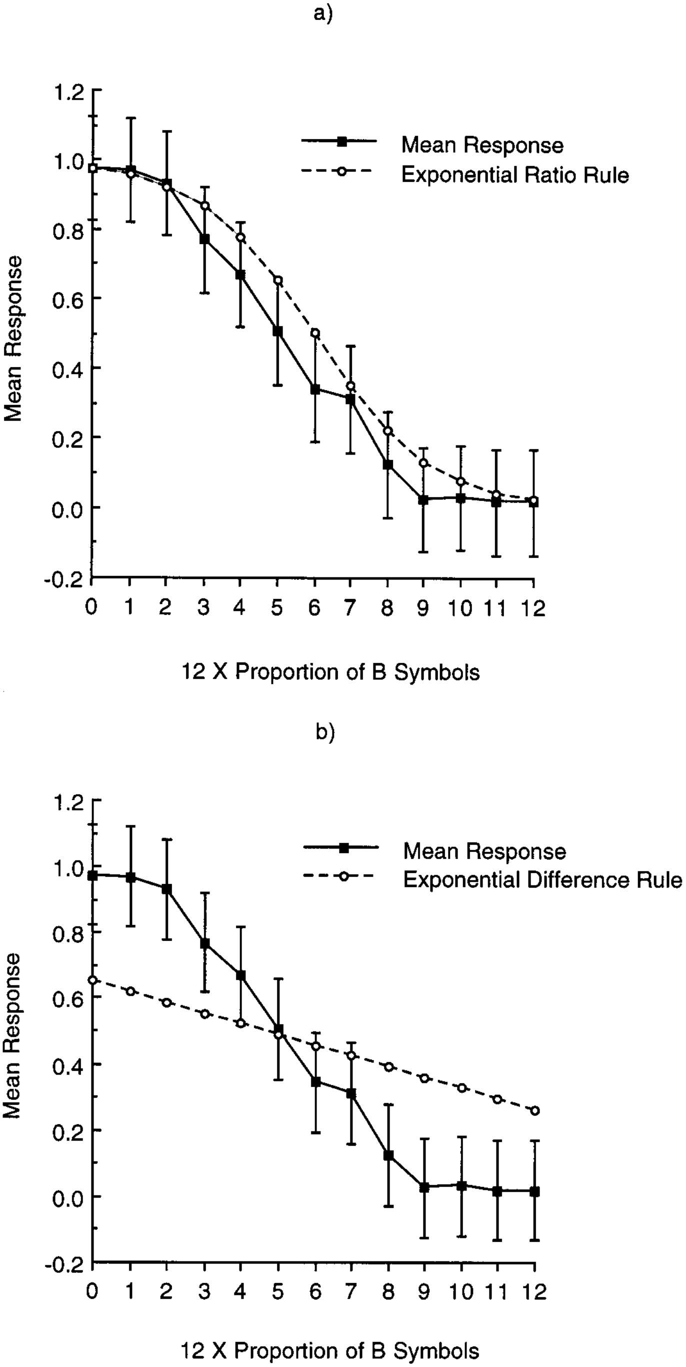

FIG. 7. A comparison between the mean response data from group $A_{30}B_{30}$/ AB* and the predictions of (a) the exponential ratio rule for $k = 4.4$. and (b) the exponential difference rule for $k = 0.8$. The error bars represent the 5% significance level and have been corrected for the number of comparisons.

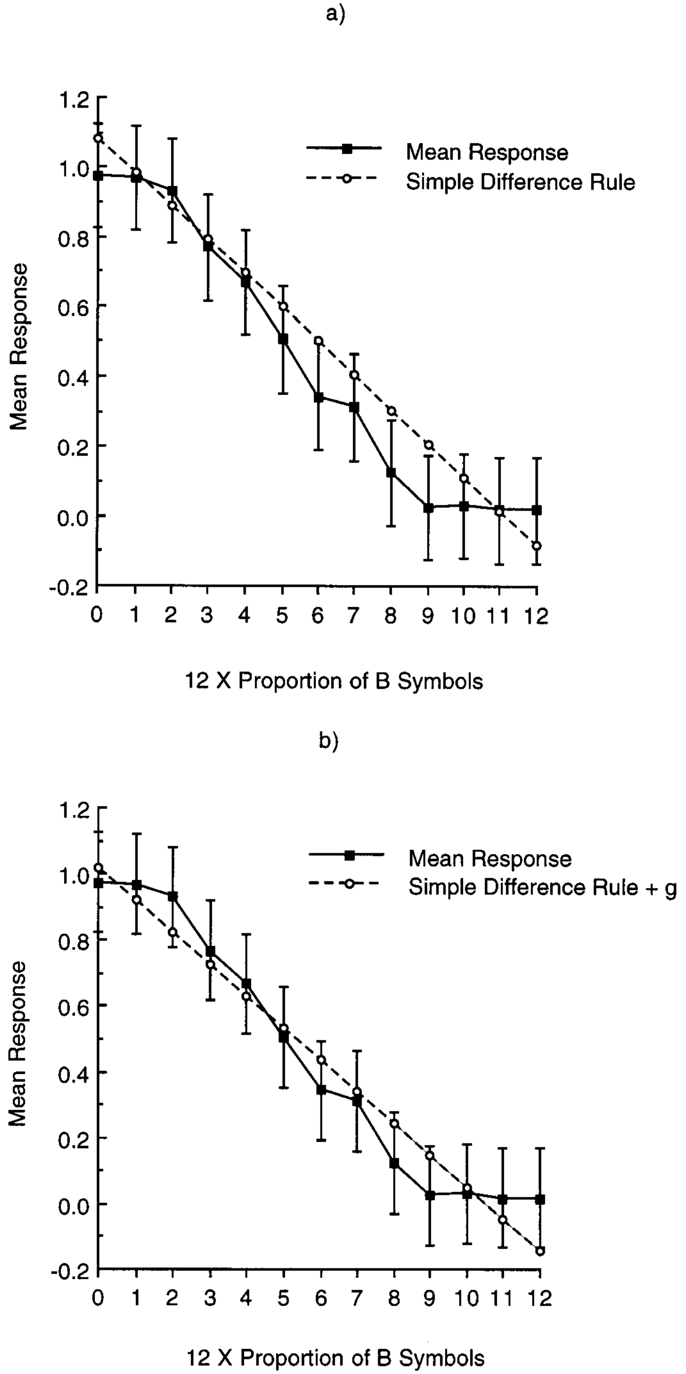

FIG 8. A comparison between the $A_{30}B_{30}$/ AB* group's mean response data and the predictions of the simple difference rule for k= 1.4 (a) in its original form, and (b) with the addition of a second free parameter (g= −0.06).

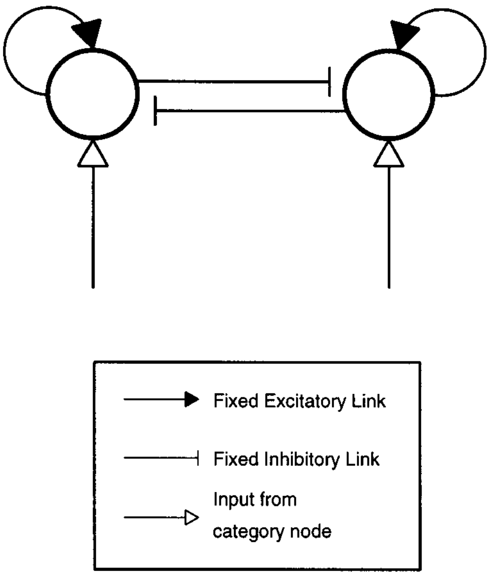

FIG. 9. The winner-take-all network architecture.

two response units competed until the activation of one exceeded the other by a specified amount. The response corresponded to the most active unit and the latency to the number of cycles it took the network to reach a decision. The probabilistic nature of the outcome was due to the presence of noise in the system.

The activation of a unit was determined using identical functions to those employed by Wills and McLaren (1997). Specifically:

$$A_c = (A_{c-1} + En) / (1 + En + D) \text{ if } n > 0 \qquad 8$$

and otherwise

$$A_c = (A_{c-1} + En) / (1 - En + \mathrm{D}) \qquad 9$$

where A_c and A_{c-1} are the activations for the current and previous cycles, respectively, E and D are constants determining the rate of excitation and decay, and n is the total input to the unit. These equations were derived from those described in McClelland and Rumelhart (1985). The units' activations were never allowed to fall below zero. n was calculated using:

$$n = e + A_{c-1} - b \qquad 10$$

where

$$e = a + \mathrm{rnd}(N) \cdot \mathrm{rnd}(1, -1) \qquad 1 \geq e \geq 0 \qquad 11$$

where a and b are the activations of the appropriate name unit and the opposite response unit, respectively, rnd(N) is a randomly selected real number between 0 and N, and rnd(1, −1) randomly produces 1 or −1. It is the last term in this equation that describes the noise in the system, and the value of N that determines its level. A new value for the noise term was calculated every cycle. Processing was terminated when the modulus of the difference between the response unit activations equalled or exceeded a threshold value, denoted by *s*. The magnitudes of *s*, *N*, *D* and *E* were 0.5, 2, 0.1 and 0.2, respectively.

The mean name unit activations that were produced by the single-layer feed-forward network, for all the groups, were input into this model. The resulting probabilities of response, which have been averaged over 1,000 simulations, are shown in Figure 10. It can be seen that the winner-take-all architecture successfully captures the empirical findings concerning mean response, at least with regard to the differences in the slopes of the generalization gradients between groups. In addition the model's predictions were compared in greater detail to the $A_{30}B_{30}$/ AB* group's mean response data, see Figure 11. This demonstrated that one point out of the thirteen lay marginally outside the 5% error bars. However, this is hardly surprising, given that no attempt was made to best fit the model's predictions to the results because of the computational difficulties involved. Indeed, the fact that even without best fitting no other points were significantly different from the data suggests that the winner-take-all network could provide at least as good an account of the mean response results as the exponential ratio rule.

A critic might argue that this is no great achievement as the model contains four free parameters. However, it should be appreciated that small changes in the values of *E*, *D* and *s* do not greatly affect the network's predictions, and that no effort was made to improve the fit by varying these paramenters.

The predictions for latencies produced by the winner-take-all architecture are shown in Figure 12. It can be seen that the model predicted a curvilinear relationship between latency and proportion of B symbols, for all groups. This is because the difference between the A and B name unit activations input into the network was greater for more extreme proportions of B symbols, resulting in a smaller number of cycles until decision for the extremes compared to intermediate proportions.

Prima facie inspection of the empirical results (Figures 2b and 5b) suggested the existence of a curvilinear relationship. In order to determine whether this was actually the case, an attempt was made to fit inverted-U shaped functions to the data from each group. This was achieved by pairing the values for the proportion of B symbols about the mid-point of 6/ 12 (e.g. 0 & 12/ 12, 2/ 12 & 10/ 12 etc.) and then, for each subject, averaging the latencies produced for these pairs. Following this transformation regression analysis revealed that only groups $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB* were adequately modelled by a straight line, which implied that the data from the other groups did not exhibit a significant U-shaped trend. However, to compensate for the possible effect of overall differences in latency between subjects, each subject's data was recoded by subtracting the mean across all proportions of B symbols from the values for each proportion. Subsequently inverted-U shaped functions provided a significant fit for all the groups.

Thus the winner-take-all model has the capacity to produce the curvilinear relationship between latency and proportion of B symbols demonstrated in the experimental results. However, the differences between groups evident in the model's predictions are not borne out by the results. This might be attributed to the high variability in the data and the fact that very little time pressure was imposed upon the subjects. Further investigation, perhaps employing tighter time constraints, would be required to establish whether the latencies generated by subjects present a significant challenge to the model.

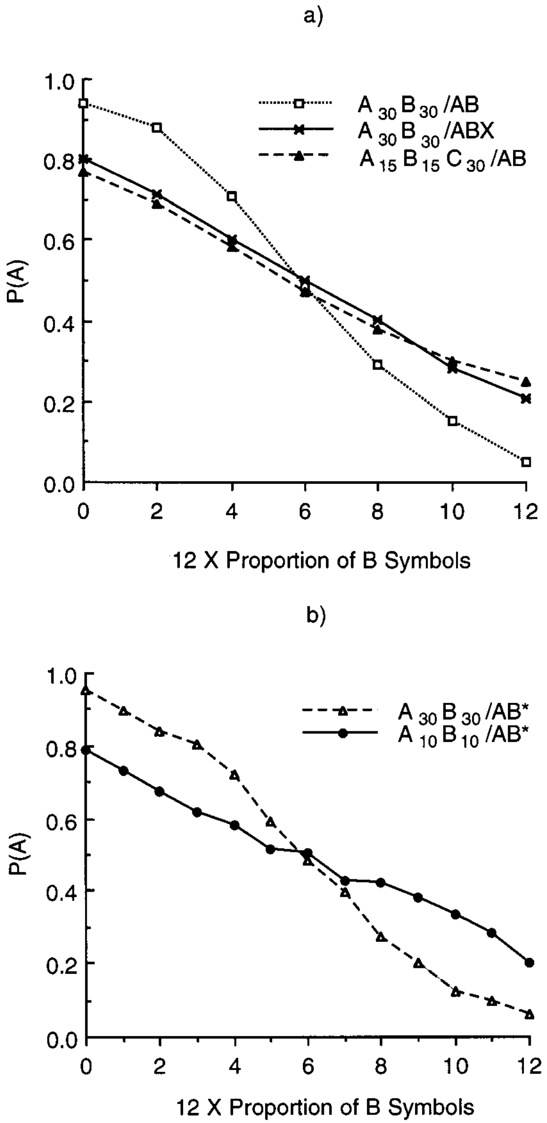

FIG. 10. The probabilities of responding "A" predicted by the winner-take-all network, averaged over 1,000 simulations, for (a) groups $A_{30}B_{30}$/ AB, $A_{30}B_{30}$/ ABX, and $A_{15}B_{15}C_{30}$/ AB, and (b) groups $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB*.

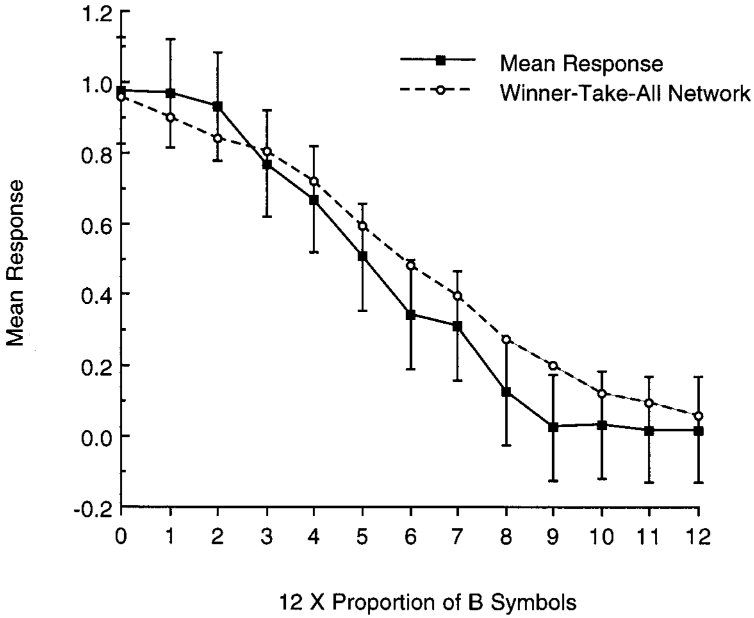

FIG. 11. A comparison between the mean response data from group $A_{30}B_{30}$/ AB*, and the predictions of the winner-take-all network. The error bars represent the 5% significance level and have been corrected for the number of comparisons.

CONCLUSIONS

Only the exponential ratio rule, operating on the activations of the name units that corresponded to responses available on test, was able to approximate to an adequate account of all the mean response data presented here. A noisy winner-take-all architecture was also found to be compatible with the mean response results. This later approach has the advantage of providing a mechanism whereby a response is made; there is something unsatisfactory about using a rule that gives a probability of response, then throwing a dice to decide the outcome. Furthermore, in contrast to many theories of categorization, this mechanism has the potential to generate latencies as well.

One additional concern might be that our analysis of decision rules only applies in the case of a simple error-correcting algorithm. We have two answers to this. The first is that there is considerable evidence for associative learning in humans and other animals being error-correcting in nature (for a review see Mackintosh, 1983; Pearce, 1997; and Shanks, 1995). The second is that, for the purposes of this research, the only important predictions of the learning algorithm are that output (category) activation is a linear function of the number of appropriate input features, and an increasing function of the length of training. As such our analysis is not directly tied to a particular learning algorithm—it applies also to any other that would result in the same input/ output function under the present experimental conditions.

We note that our analysis and, in particular, the decision network, can equally be applied to performance in animals other than humans. As an example that draws on

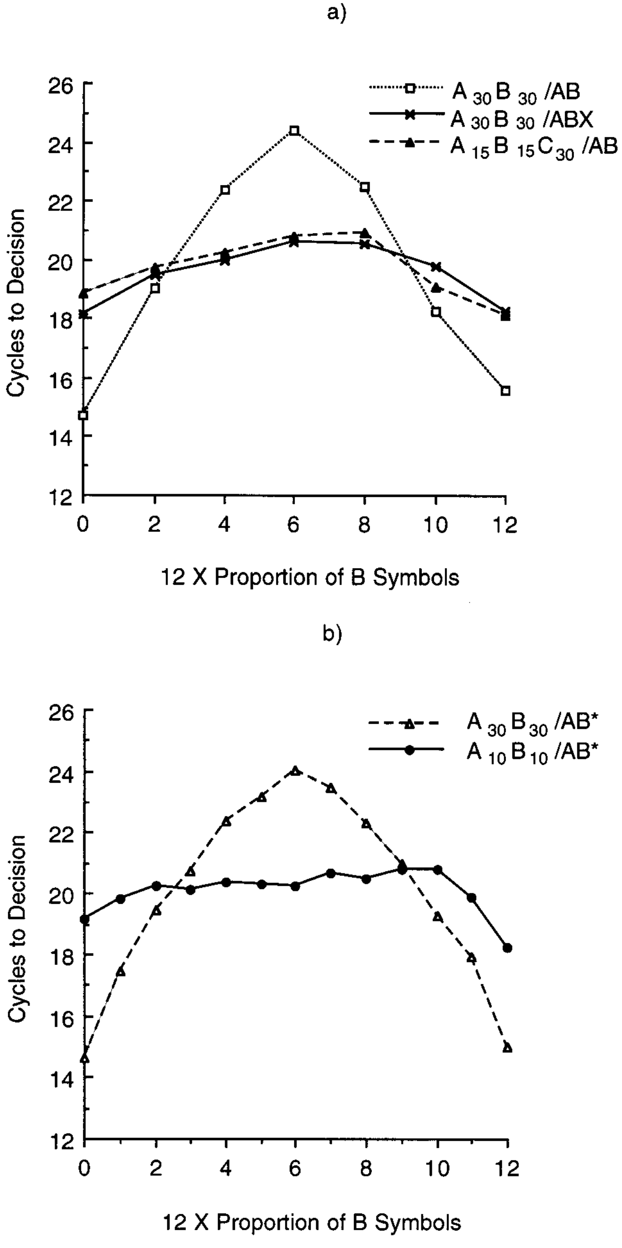

FIG. 12. The mean number of cycles for the winner-take-all network to reach a decision, averaged over 1,000 simulations, for (a) groups $A_{30}B_{30}$/ AB, $A_{30}B_{30}$/ ABX, and $A_{15}B_{15}C_{30}$/ AB, and (b) groups $A_{30}B_{30}$/ AB* and $A_{10}B_{10}$/ AB*.

the studies reported here, consider the issue of summation. If stimuli X and Y are reinforced and then tested in compound as XY, should we expect a higher rate of responding to the compound than to X or Y? Our answer would be that you might get a higher rate of responding (summation) after relatively little training, but little evidence for summation after more prolonged training that took performance on the training discriminations nearer asymptote, because at this point increments in associative strength will have little effect on response probability. This follows from the assumption that, under these circumstances, subjects will be operating in the relatively flat region of the function relating activation (or associative strength) to response probability (or response rate) for the stimuli in question.

The foregoing example makes it clear that any failure to observe summation as predicted by a Rescorla–Wagner (1972) type analysis need not be taken as evidence against that learning algorithm. In this light, it is not surprising that if X and Y are two stimuli relatively close to one another on some dimension and testing is to some intermediate value on this dimension, then better responding to the test stimulus than to either X or Y is not observed (Mackintosh, 1974, pp. 532). Once again we would argue that although the associative strength from the test stimulus to the US might well be greater than that for X or Y, the flat response function at this level of associative strength prevents this difference from manifesting. On the other hand, the result that a form of summation can be observed when two reinforced stimuli are relatively far apart on some dimension and responding to an intermediate test stimulus is assessed fits well with the response function generated by the decision mechanism offered here.

We believe that the further investigation of tasks involving more than two categories will provide an interesting test for both the winner-take-all mechanism and the exponential ratio decision rule. If, as we suspect, the winner-take-all network can account for the data from studies of this type and its predictions concerning latencies are borne out, then it has the potential to be applied far more widely than just to simple discriminations. Indeed, it could, in principle, be used to model any situation in which the output of a connectionist network needs to be converted to one of a particular set of competing responses.

REFERENCES

Bradley, R.A. (1954). Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparison. *Biometrics*, *10*, 375–390.

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

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

Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Hopkins, J.W. (1954). Incomplete block rank analysis: Some taste test results. *Biometrics*, *10*, 391–399.

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of mathematicl psychology: Volume 1*. New York: Wiley.

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

Mackintosh, N.J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, *88*, 375–407.

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

McLaren, I.P.L., Bennett, C.H., Guttman-Nahir, T., Kim, K., & Mackintosh, N.J. (1995). Prototype effects and peak shift in categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 662–673.

Pearce, J.M. (1997). *Animal learning and cognition. An introduction* (2nd ed.). Hove, East Sussex: Psychology Press.

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: Vol. II. Current research and theory*. New York: Appleton-Century-Crofts.

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

Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Wagner, A.R., Logan, F.A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76*, 171–180.

Wills, A.J., & McLaren, I.P.L. (1997). Generalization in human category learning: A connectionist account of differences in gradient after discriminative and non-discriminative training. *Quarterly Journal of Experimental Psychology, 50A*, 607–630.

Manuscript received 26 June 1997
Accepted revision received 15 October 1997

Catégorisation perceptuelle: modèles connexionistes et règles de décision

Même si les modèles de l’apprentissage associatif chez les humains en termes de réseaux connexionistes sont populaire, le mécanisme par lequel les activations terminales (<<output activations>>) sont converties en probabilités de réponse a reçu très peu d’attention. Plusieurs modèles de ce processus décisionel sont considérés incluant une simple règle de proportion, une simple règle de différence, leurs versions exponentielles, et un réseau gagnant-prend-tout (<<winner-take-all network>>). Deux expériences de catégorisation qui ont tenté de séparer ces modèles sont rapportées. Des analogues de ces expériences furent présentés à un simple réseau connexioniste utilisant l’algorithme d’apprentissage delta. Seulement la version exponentielle de la règle de proportion et l’architecture gagnant-prend-tout, opérant sur les activations terminales des réseaux connexionistes qui correspondent aux réponses durant le test, furent capable de prévoir de façon compréhensive les résultats. De plus, à l’encontre de la version exponentielle de la règle de proportion, le modèle gagnant-prend-tout peut, potentiellement, prévoir les temps de réponses. Des recherches futures vont devoir déterminer si ces temps de réponses produits dans des conditions plus rigoureuses se conforment aux prédictions de ce modèle.

Categorización perceptiva: modelos conexionistas y reglas de decisión

Aunque actualmente con frecuencia se modela el aprendizaje asociativo humano usando redes conexionistas, el mecanismo por el que sus activacíones de salida se convierten en probabilidades de respuesta ha recibido relativamente poca atención. Se consideran diversos posibles modelos de este proceso de decision incluyendo una regla de razón simple, una regla de diferencia simple, sus versiónes exponenciales y una red todo-para-el-ganador. Se llevaron a cabo dos experimentos de categorización que intentaron separar estos modelos. Se presentaron análogos de los experimentos a una red de una sola capa, regla delta y propagación hacía adelante. Solo la regla de razón exponencial y la arquitectura todo-para-el-ganador, actuando sobre las activacíones de salida de la red que correspondian a respuestas disponibles sobre la prueba, fueron capáces de predecir completamente los resultados de la media de respuesta. Además, a diferencia de la regla de razón exponencial, el modelo todo-para-el-ganador tiene la potencialidad de predecir latencias. Seran necesarios nuevos estudios para determinar si las latencias producidas bajo condicíones mas rigurosas se ajustan a las prediccíones del modelo.