A Connectionist Model of Category Size Effects During Learning

Timothy J. Breen
Boeing Advanced Technology Center

ABSTRACT

This paper reports the results of category learning experiments in which the number of exemplars defining a category during learning was varied. These results reveal that category exemplars from larger sized categories are classified more accurately than those from smaller-sized categories. This was true both early and late in learning. In addition, subjects exhibited a response bias toward classifying exemplars into larger-sized categories throughout learning. A connectionist model is developed which exhibits these same tendencies.

INTRODUCTION

This paper reports the results of category learning experiments in which the number of exemplars defining a category during learning was varied. These results are then compared with the results of simulations using a connectionist model of category learning. Categorization has a special status for connectionist models, since the ability of connectionist systems to learn generalizations from specific instances is frequently cited as one of the most promising aspects of the connectionist approach (e.g. Norman, 1986, pp. 535-536). Although several examples exist in which connectionist models have been successfully applied to data from classification experiments (e.g. Knapp & Anderson, 1984; McClelland & Rumelhart, 1985; Gluck & Bower, 1988) the range of these cases is relatively narrow. Therefore, it is important to evaluate these models in light of additional empirical findings.

THE EFFECT OF CATEGORY SIZE ON LEARNING RATE

A robust finding in the classification literature is that increasing the number of exemplars representing a category during learning, under most circumstances, improves transfer performance on novel category exemplars (e.g. Homa & Vosburgh, 1976). What is not known is whether, or how, this variable influences category learning. For example, in a task in which subjects are required to learn the category assignments of members of three different categories, where the categories contain 3, 6, and 9 members respectively, is one category learned more quickly than the others? One might suspect for example, that the category with only three members would be easiest for subjects to learn.

EMPIRICAL FINDINGS

To examine this question, analyses of previously unreported data from a series of experiments conducted by Breen & Schvaneveldt (1986) are reported below. Breen & Schvaneveldt conducted three experiments in which subjects learned to classify dot patterns (Posner, Goldsmith & Welton, 1967) into three different categories. Dot pattern categories have been used extensively in the classification literature, and are constructed by first assigning dots (usually nine) randomly into cells of a matrix. This dot pattern is referred to the objective prototype of the category. To generate category exemplars, a statistical distortion rule is applied to the objective prototype that moves the dots to a new position in the matrix. Additional categories can be created by generating distortions of a new random objective prototype pattern.

In the these experiments, categories in the learning phase were represented by 3, 6, or 9 dot patterns. Subjects continued to classify patterns during the learning phase until all 18 patterns were

classified correctly during a single block of trials. Conditions in the learning phase for all three experiments were identical, and because transfer performance was of primary interest, the learning data were not reported earlier. (See Breen & Schvaneveldt, 1986, for further details of the experimental procedure).

Out of 300 subjects participating in all three experiments, 44 failed to reach learning criterion (errorless performance in 30 blocks or less in Experiment 1, or 35 blocks or less in Experiments 2 & 3), and these data were excluded from the analyses. The average number of blocks to criterion for all remaining subjects was 15.5.

Figure 1 shows average correct responses over learning blocks for each of the three category sizes. To generate these learning functions, errorless performance was assumed after each subject achieved the learning criterion. For example, if a particular subject met the learning criterion after 10 blocks of trials, it was assumed that no errors would have occurred for blocks 11 through 35. Since this assumption is probably too strong, the right-hand side of the graph in Figure 1 is most likely artificially inflated for all category sizes. It is clear, however, that early in learning, classification accuracy was enhanced for exemplars of the larger categories.

Figure 2 shows a clearer picture of classification accuracy late in learning, in which classification accuracy is plotted as a function of the $-n^{th}$ block in relation to each subjects' learning criterion (backward learning curve, Trabrasso & Bower, 1968). The number of subjects contributing to each data point is also shown in the bottom of the figure. Surprisingly, Figure 2 suggests that the larger category sizes maintained their advantage late in learning.

To confirm these results, an analysis of variance was performed on data from the first three blocks of trials and for the last three prior to reaching the learning criterion for each subject. Because 46 subjects reached criterion in less than seven blocks, these data had to be excluded from this analysis. The analysis treated category size as a factor with three levels (3, 6, and 9), and blocks as a factor with two levels (early and late) in a (3 x 2) factorial design. The results revealed a main effect of blocks [$F(1,209) = 687.87$, $MSe = 29.881$, $p<.001$], and category size [$F(2,418) = 24.03$, $MSe = 0.791$, $p<.001$]. Category size did not interact with blocks [$F(2,418)=0.34$,

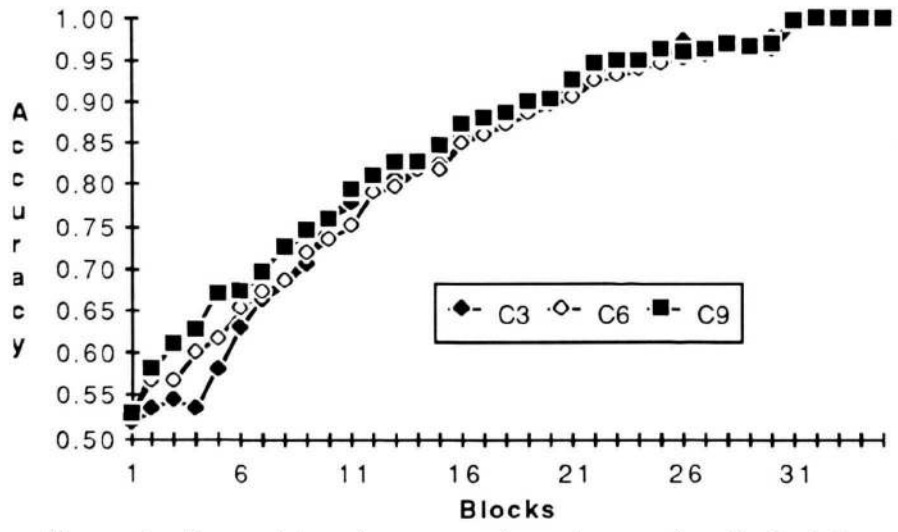


Figure 1. Forward learning curves for category sizes 3, 6, & 9.

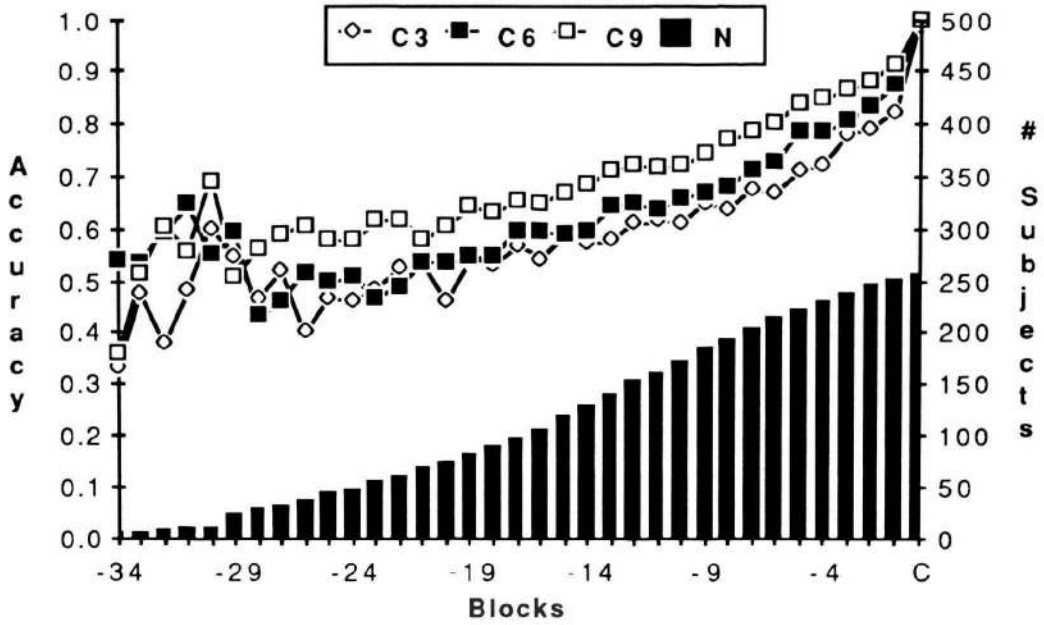


Figure 2. Backward learning curves for category sizes 3, 6, and 9.

MSe=0.006], suggesting that the effect of category size was the same both early and late in learning.

To summarize, these results suggest that those categories containing a larger number of exemplars were learned most quickly, and that subjects were more accurate in classifying exemplars from large categories both early and late in learning. These findings are problematic for at least some distributed models of learning and memory.

McCLELLAND & RUMELHART'S (1985) MODEL

For example, McClelland and Rumelhart (1985) have proposed a model of category learning and representation that employs the delta learning rule (Figure 3). Since this model has been described in detail elsewhere, only a brief description of the general properties of the model will be presented here. The model consists of a single layer of nodes, with each node in the model connected to every other node. Each node may receive activation from two sources. One is from outside the network when a pattern, in the form of a binary feature vector, is presented to the model. The other is from other nodes in the network through connections which have non-zero weights.

The model is trained by presenting a pattern to the model, allowing activation to spread throughout the nodes, and then applying the delta rule to adjust the connection weights such that the activity levels of the nodes match, or come progressively closer to matching, the input pattern. The delta learning rule is specified by:

$$W_n = W_{n-1} + \eta \delta_n \, i^T{}_n$$

where W_n is the weight matrix following trial n, η is a constant which determines the rate of learning, $i^T{}_n$ is the transpose of the input pattern on trial n, and δ_n is the difference between the desired and actual output on trial n:

$$\delta_n = t_n - W_{n-1} \, i_n$$

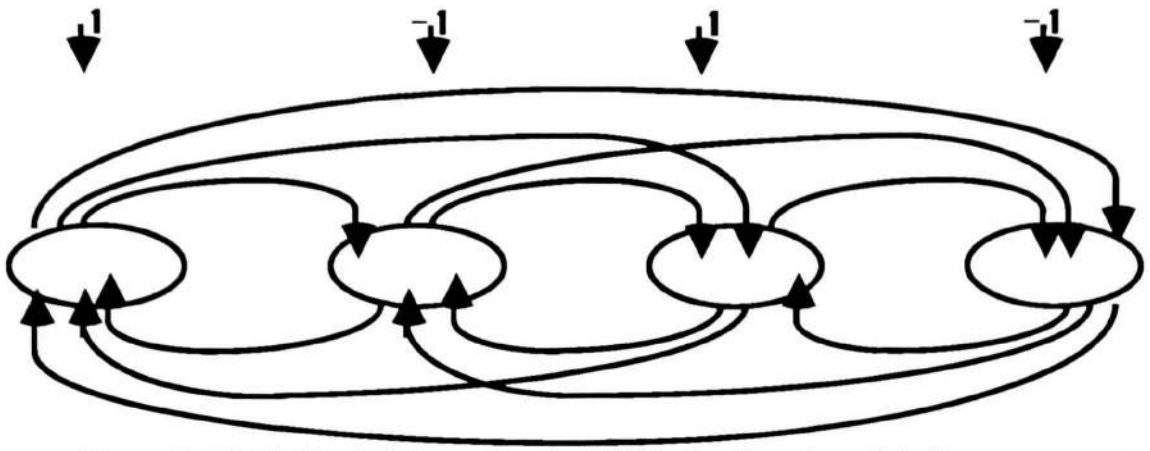


Figure 3. McClelland & Rumelhart's (1985) distributed model of memory.

where t_n is the desired or target output on trial n and $W_{n-1}\, i_n$ is the actual output produced on trial n. In the McClelland & Rumelhart (M&R) model, the target output value t in the above equation is the input pattern on a particular learning trial.

The performance of the model is evaluated in terms of the hacking distance between the input pattern and resulting node activations. This measure is referred to as *response strength*. The general idea is that when the pattern of activation produced by the model closely matches the input pattern, the input pattern has closely matched what is stored in the connection weights. In other words, if response strength is high, the model has recognized the input as something that it has learned or knows about. The response strength for input pattern p is the dot product over the activations of each node and the input pattern, normalized for the number of nodes in the model:

$$RS_p = \frac{1}{n} \sum_{i=0}^{i=n} a_i p_i$$

A SIMULATION

The ability of the M&R model to account for the above results is evaluated in the following simulation. For the simulation, training patterns from different categories were constructed by first generating three random binary feature vectors of length 20. These patterns become the category objective prototypes. Distortions of the objective prototypes were then generated by flipping the sign of each feature in the objective prototype with a probability of .15. The training set consisted of 3 distortions of one prototype, 6 distortions of a second, and 9 distortions of a third, for a total of 18 patterns.

During each trial in the simulation, a training pattern was presented to a model consisting of 20 completely connected nodes, activation was allowed to spread and stabilize throughout the model, then the connection weights were changed according to the delta learning rule. Each block of trials consisted of one pass through the 18 patterns, and each simulation run consisted of 30 blocks of trials. The number of simulation runs, consisting of stimulus generation-model training cycles, was 100.

Figure 4 plots response strength of the model over learning blocks. Average response strength is greater for members of the category containing nine patterns early in learning, but average response

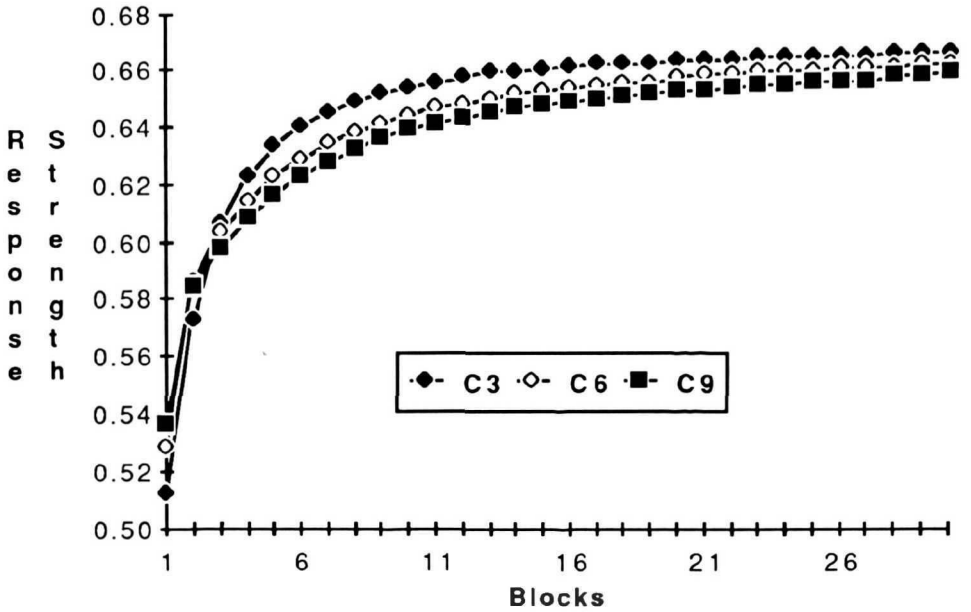


Figure 4. Response strength plotted over 30 learning blocks for category sizes 3, 6, &9.

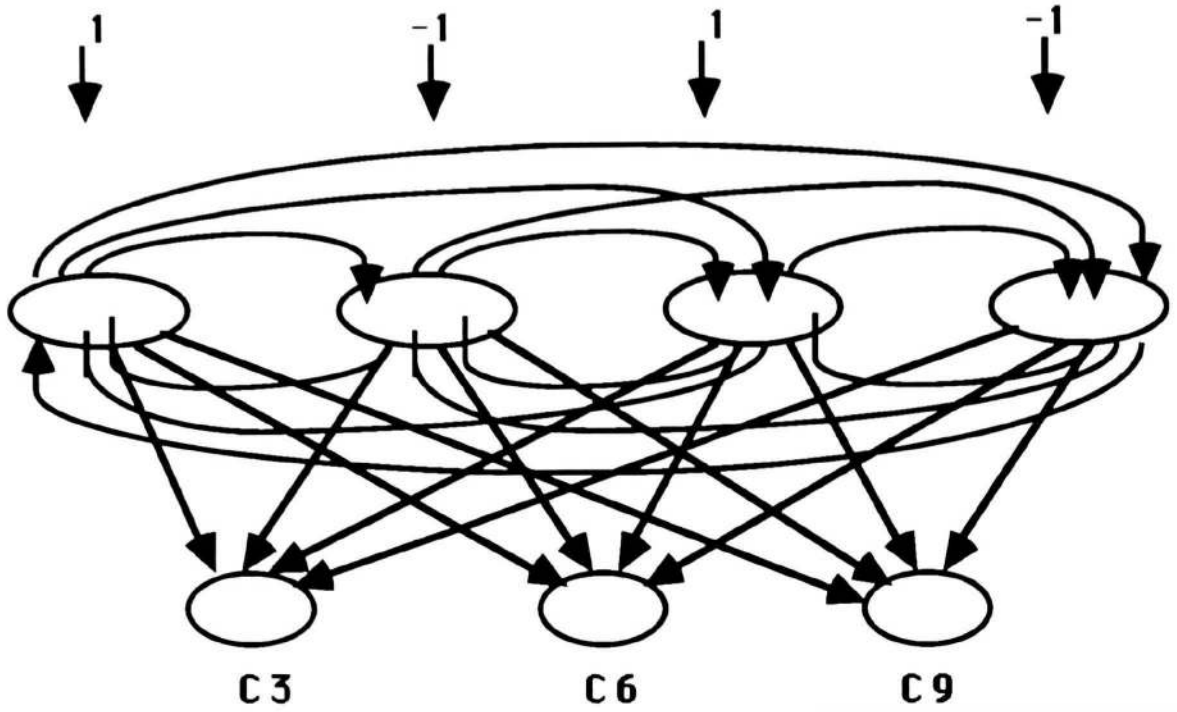
strength for members of the category containing three members quickly overtakes response strength for other category members.

Considering the general properties of the model provides some insight into the simulation results. The model has the ability to represent both general, abstract information, along with specific, instance information in the same connection weights. In this sense, it is similar to a mixed-prototype model of categorization (e.g. Homa, Sterling, & Trepel, 1981). One factor which determines how strongly the model represents general or specific information about a category is how many distinct patterns comprise a category during learning. In general, the model retains highly specific information about small categories, and more abstract information about large categories. Under most circumstances, this results in more accurate generalizations to novel patterns when trained on greater numbers of distinct category exemplars (Breen, 1988).

The interaction shown in Figure 4 is made clear by considering that on each block of learning trials, half of the patterns belonged to the largest category. This caused the early advantage for the category with 9 members, because the model had relatively more experience with that category. Why the slope of the learning function is steepest for the smallest category is precisely because there were only three patterns to learn. That is, more interference among same category members is expected to occur as category size increases, producing a flatter learning function. This property of the model instantiates the mixed-prototype model assumption that processing capacity limitations (among other things) encourage abstract representations.

AN EXTENSION OF THE MODEL

A simple extension of the M&R model would involve the addition of a set of output nodes, with each output node responding to evidence concerning the presence of a particular category. In this model, shown in Figure 5, the input layer is completely connected and is trained the same way as before, by using the delta rule to produce a pattern of activity across the nodes that matches the input pattern. In addition, each node in the input layer is connected to each node in the output layer. The delta rule is also used to train the output nodes to produce a pattern of activity



Category	Desired Response
C3	1 0 0
C6	0 1 0
C9	0 0 1

Figure 5. An extension of McClelland & Rumelhart's (1985) model.

that more closely resembles a categorization response. For example, consider the previous simulation in which the model is trained on patterns from three categories, with each category containing either three, six, or nine exemplars during learning. Each node in the output layer can be trained to take on positive activation depending on which category C3, C6, or C9, an input pattern belongs. For example, if a pattern from C3 (the category containing 3 exemplars) is presented to the model, the output layer is trained to produce the activity pattern [1 0 0] (see Figure 5).

The previous simulation in which category size was varied during learning was repeated using the model in Figure 5 (referred to as Model 2). All other methodological aspects of the simulation were identical to the method employed earlier. The sequence of events on each learning trial was as follows. A pattern was presented to Model 2, and activation was allowed to spread throughout the network (both input and output layers) until these activation levels stabilized. The activity levels in the input layer were then matched against the input pattern, and the weights connecting nodes in the input layer were adjusted using the delta rule. Simultaneously, the activity pattern in the output layer was compared to the desired category response, which is shown in Figure 5 for each category, with the delta rule again determining weight adjustments from the input to output layers.

Figure 5 shows the activity levels of nodes in the output layer for "correct" category nodes, for example, the average activity level for node C6 when a pattern from the category containing 6

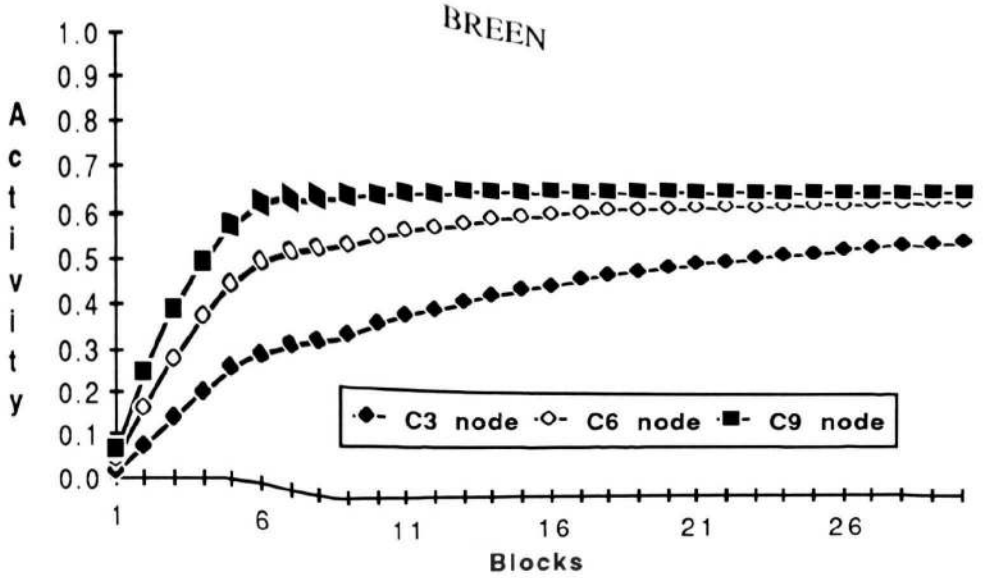


Figure 6. Average activity levels for "correct category" nodes.

exemplars was presented. The results of this simulation show that the activity level in the output nodes corresponding to the larger categories remained consistently higher than the activity levels of the nodes corresponding the smaller sized categories. This can be seen most clearly by comparing the activity levels of the C3 and C9 nodes in Figure 5. The results of this simulation are more in line with the results of the Breen & Schvaneveldt experiments.

Recall that one of the reasons cited for the inability of the M&R model to account for this result was that storing a large number of exemplars from the same category in the connection weights tends to produce interference in the input layer, producing a flatter learning function. This interference, however, only concerns the ability of the model to respond strongly to specific (old) input patterns. More exemplar experience also produces more accurate generalizations With increased experience, what the model gives up in representing specific information it gains in representing generality. Interference, per se, is thus not an undesirable quality. The same holds true for the input layer in Model 2. However, because the output layer of Model 2 is trained to produce a category level response, increased training on different patterns from the same category will only facilitate the acquisition of category-level information by the model.[1]

RESPONSE BIASES

Two further questions can be addressed by an analysis of the Breen & Schvaneveldt learning data that involve the particular kinds of errors that subjects make while learning to classify exemplars of categories which vary in size. The first question is whether category size influences the kinds of errors subjects make during learning. For example, when an error is made when classifying an exemplar from a category of size six, are subjects more likely to classify it as a member of the larger (size nine) category? This would be expected if subjects are using information about the relative size or likelihood of the three categories in making a response. The second question is that if subjects are prone to a response bias of this nature, will this bias be differentially reflected in errors occurring early and late in learning? One possibility is that such a bias would more strongly

[1]The connections between nodes in the input layer do not play a role in accounting for this category size effect. For example, an independant-cue model of the type proposed by Gluck and Bower (1988) is able to produce this same behavior, as well as the "response bias" tendencies in the following section.

influence responses early in learning, when subjects have less complete knowledge about category membership. For instance, when subjects are unsure of the correct category assignment when an exemplar is presented, they may base their response on knowledge about the relative probability of category exemplars occurring on each trial. And, this may more frequently occur early in learning, before much category information has been acquired.

EMPIRICAL FINDINGS

Figure 7 shows the breakdown of errors occurring during the first three learning trials (Early) and the last three trials before criterion (Late) for 210 subjects. It shows that when an exemplar from one of the three categories (C3, C6, or C9) was presented during learning, subjects were more likely to make an error by classifying the exemplar into a relatively larger sized category. In addition, this trend is equally apparent both early and late in learning. The magnitude of the bias appears to be greatest when a member from C6 is presented. This is consistent with the explanation that subjects were using probability information about the relative frequency of occurrence of category exemplars during learning, since C9 and C3 are the largest and smallest categories.

To confirm these results, an analysis of variance was performed treating Blocks as a factor with two levels (early and late), Response as a factor with three levels (C3, C6, and C9),and Correct Category (or category size) as a factor with three levels (C3, C6, and C9). In addition to the main effects reported above, this analysis revealed a main effect of Response [$F(1,209) = 31.010$, $MSe = 0.937$, $p<.001$]. The Response by Correct Category interaction approached significance [$F(2,418) = 2.797$, $MSe = 0.076$, $p<.1$], as did the three-way interaction of Blocks, Correct Category, and Response [$F(2,418) = 2.620$, $MSe = 0.041$, $p<.1$]. Blocks and Response did not interact [$F(1,209) = 1.935$, $MSe = 0.028$, $p>.1$].

It appears that subjects were prone to bias their responses toward the larger-sized categories to the same degree both early and late in learning. The finding that Blocks and Response did not interact was somewhat surprising, because it might be expected that a response bias would be reflected to a greater degree during the early blocks, when category learning is minimal. However, the acquisition of knowledge relating to the category membership of particular exemplars was confounded with the acquisition of knowledge about the relative sizes of each category in these experiments. Subjects were not told prior to the experiment that each category was represented by a different number of members during the learning phase. So early in learning, category size information may have been available to a lesser degree relative to later stages in learning.

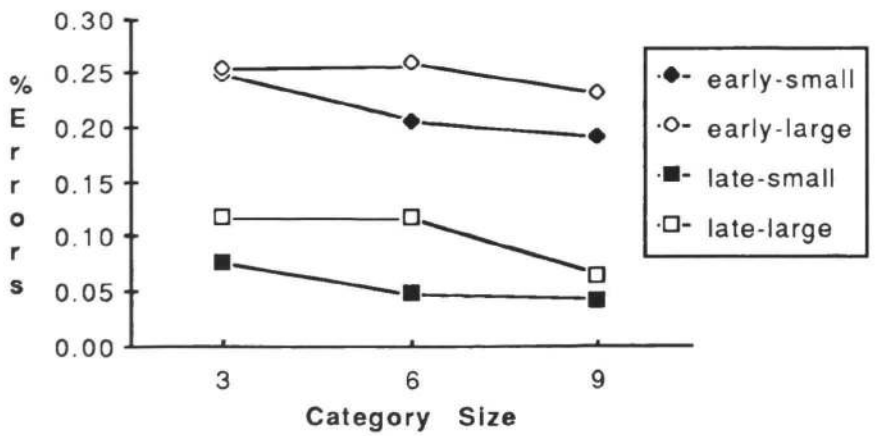


Figure 7. P(error) for first and last three blocks as a function of response and category size.

Therefore, a model that proposes that frequency information plays a stronger role in the absence of more "categorical" knowledge may still be consistent with these data. In these experiments, such a model would assume that with more experience in classifying exemplars during learning, the quality of both frequency and category information is enhanced. Early in learning, subjects rely to a relatively greater extent on poor quality frequency information. And late in learning, subjects rely to a lesser degree on high quality frequency information.

The above discussion, of course, lacks a connectionist flavor. Any model incorporating the notion of a response bias, which seems most naturally described in terms of rules or strategies, is inconsistent with the spirit of connectionist modeling. Ideally, a connectionist model's behavior should exhibit a tendency toward classification into larger sized categories and arise naturally from the structure of the input population and the architecture of the model.

SIMULATION RESULTS

The potential ability of Model 2 to account for these results can be examined in a straight-forward manner by observing the model's performance during the previous simulation. In particular, when a pattern from a particular category is presented during learning we can observe the activity levels in the nodes corresponding to the incorrect categories. For example, when a pattern from the category containing six members is presented (C6) to the model during learning, what are the activity levels of nodes corresponding to C3 and C9? Figure 8 shows these values across 30 learning blocks during the previous simulation.

Figure 8 shows that when Model 2 was learning to classify patterns from three categories containing either three, six, or nine patterns, and was presented with a pattern from the category containing six patterns, the activation of the C9 node was consistently higher than the activation of the C3 node. In fact, during learning, the model showed a general tendency to slightly inhibit those nodes corresponding to the two alternative categories, and the degree of inhibition depended upon category size. Nodes corresponding to smaller categories were inhibited to a greater extent than larger categories on those trials when an alternative category pattern was presented. This is somewhat interesting behavior from a model that contains no explicit mechanisms for producing a "response bias" for larger sized categories.

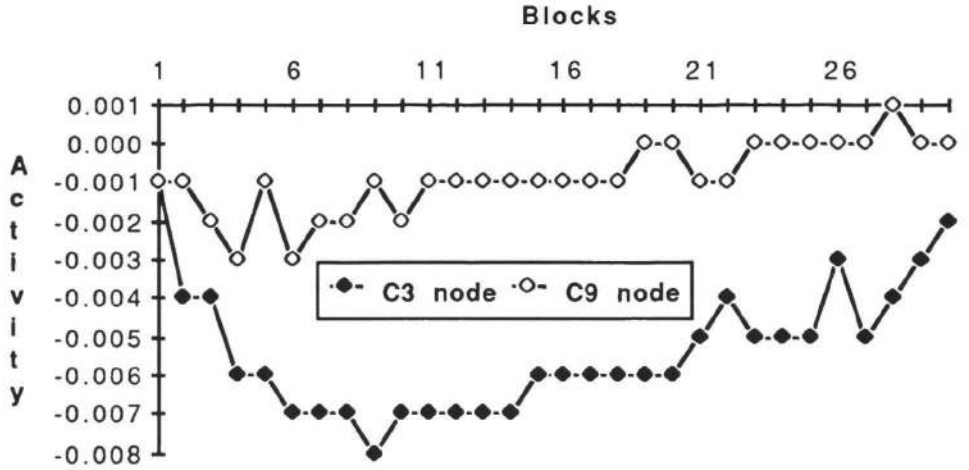


Figure 8. Activity levels of C3 and C9 node during learning trials when C6 pattern was presented.

Although the model as it stands is clearly too underdeveloped to make quantitative predictions about subjects behavior in this task, the model does exhibit a completely natural tendency toward inhibiting classification into relatively smaller-size categories. One further note is that when the learning procedure involves actively inhibiting alternate category responses, it will produce radically different behavior. For example, if on a particular trial the output layer is trained to produce the activity pattern [-1 1 -1] instead of [0 1 0] when presented with a pattern from C6, the model will learn to more strongly inhibit the C9 node, which produces response bias in the opposite direction than before. This finding produces a further constraint on the particulars of the learning procedure.

CONCLUSIONS

An extension of McClelland & Rumelhart's (1985) distributed model of learning and memory was shown to account (at least qualitatively) for subjects behavior in a category learning task in which category size was varied. Other researchers, no doubt, will fault the model for its inherent linearity. However, linear models have been found to be surprisingly robust over a variety of conditions in simulations of categorization tasks (Breen, 1988). All models can be pushed past their limit, and the present work is intended to provide some useful constraints for further model development.

REFERENCES

Breen, T. J. (1988). *An Evaluation of Connectionist Models of Categorization.* Unpublished doctoral dissertation. New Mexico State University, Las Cruces, New Mexico.

Breen, T. J., & Schvaneveldt, R. W. (1986). Classification of empirically derived prototypes as a function of category experience. *Memory &Cognition*, **4**, 313-320.

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

Homa, D., Sterling, S. & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 418-439.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, **2**, 322-330.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 616-637.

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.

Norman, D. A. (1986). Reflections on cognition and parallel distributed processing. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition, (Vol. 2: Psychological and Biological Models).* Cambridge, Mass.: MIT Press.

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, **77**, 353-363.

Trabasso & Bower, G. H. (1968). Attention in learning: Theory and research. New York: Wiley.

ACKNOWLEDGMENTS

This paper is based on portions of a PhD dissertation in psychology submitted to New Mexico State University. The research was supported by the Computing Research Laboratory at NMSU. I would like to thank Roger Schvaneveldt, Jim McDonald, Jordan Pollack, Ken Paap and Don Dearholt for many contributions during the dissertation work, and Mark Gluck, Keith Butler, and Colleen O'Neill for critical reading of an earlier version of this paper.