Research Report

# STIMULUS GENERALIZATION AND REPRESENTATION IN ADAPTIVE NETWORK MODELS OF CATEGORY LEARNING

Mark A. Gluck
*Stanford University and Rutgers University*

**Abstract**—*An exponential-decay relationship between the probability of generalization and psychological distance has received considerable support from studies of stimulus generalization (Shepard, 1958) and categorization (Nosofsky, 1984). It is shown here how an approximate exponential generalization gradient emerges from stimulus representation assumptions isomorphic to a special case of Shepard's (1987) theory of stimulus generalization in a "configural-cue" network model of human learning that represents stimulus patterns in terms of elementary features and pair-wise conjunctions of features (Gluck & Bower, 1988b; Gluck, Bower, & Hee, 1989). The network model can be viewed as a combination of Shepard's theory and an associative learning rule derived from Rescorla and Wagner's (1972) theory of classical conditioning.*

A well-established principle of stimulus generalization is the approximate exponential-decay relationship between the probability of generalization and psychological distance (Shepard, 1958, 1987). This principle has received additional support from Medin and Schaffer's (1978) exemplar-context theory of classification. In their theory, a test pattern acts as a retrieval cue to access information associated with similar stored exemplars. The similarity between two exemplars is computed according to a rule that assumes similarity decays exponentially with increasing distance in an appropriate psychological space (Nosofsky, 1984). This model has had considerable success in accounting for many human classification and recognition behaviors (Estes, 1986; Smith & Medin, 1981).

Correspondence and reprint requests to Mark A. Gluck, Department of Psychology, Jordan Hall; Bldg. 420, Stanford University, Stanford, CA 94305; or electronic mail to gluck@psych.stanford.edu.

An alternative class of models, based on adaptive "connectionist" networks, is also able to account for many of these behaviors (Gluck, Corter, & Bower, 1990; Gluck & Bower, 1988a, 1988b; Gluck, Bower, & Hee, 1989, 1990). The network model, shown in Figure 1, adapts its weights (associations) according to Rescorla and Wagner's (1972) model of classical conditioning. The presentation of a stimulus pattern is represented by activation of nodes on the input layer that correspond to the pattern's elementary features and pair-wise conjunctions of features. The inclusion of conjunctive cues allows the model to solve complex discriminations that require sensitivity to conjunctions of features; such discriminations are called "nonlinearly separable." This model is formally equivalent to Wagner and Rescorla's (1972) "configural-cue" proposal for extending their conditioning model; studies of animal learning have found considerable predictive and explanatory power for this model (Kehoe & Gormezano, 1980; Rescorla, 1972, 1973). It is also a special case of "higher-order" networks (also called "polynomial" or "functional link" networks) that have been used by adaptive network theorists as an alternative to multilayer networks (Pao, 1989).

Like the exemplar-context model,

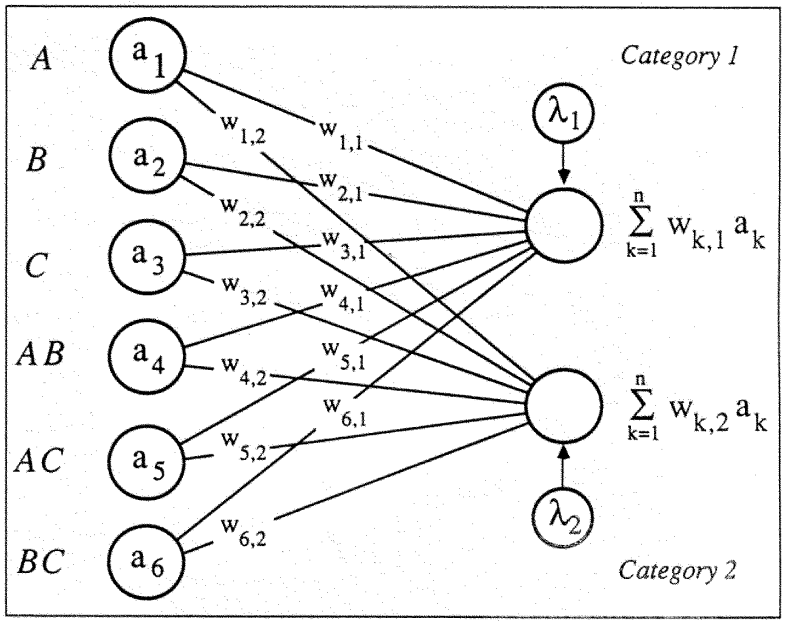


Fig. 1. A configural-cue network model of human learning that represents stimulus patterns in terms of their elementary features and pair-wise conjunctions of features. In this example, input nodes code for the presence or absence of three component-cues (A, B, C) and all possible pair-wise configural combinations of these cues (AB, AC, BC). The network's classification prediction is a function of the resulting activation on the output nodes. Associative weights between feature and category nodes are updated according to the error-correcting principle of the Rescorla–Wagner (1972) model of classical conditioning, equivalent in this application to Widrow and Hoff's (1960) LMS rule of adaptive network theory.

  

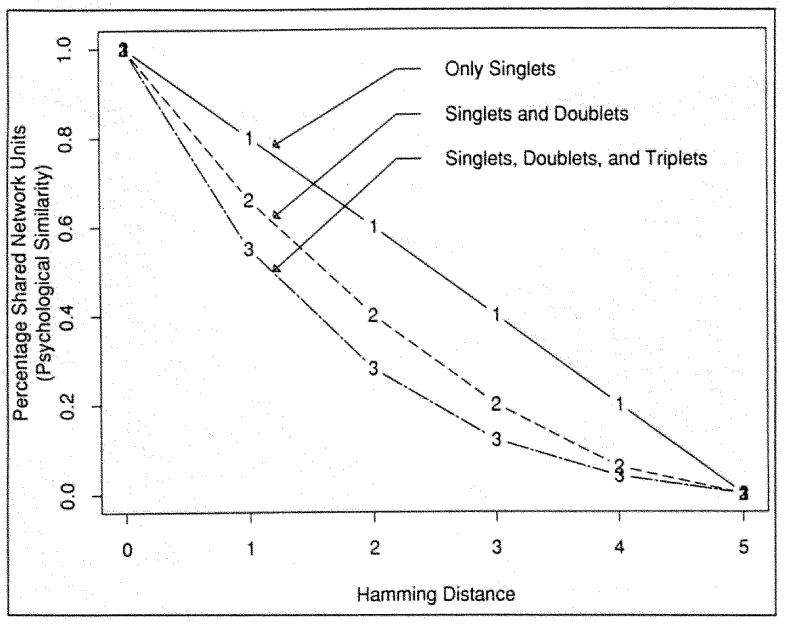


**Fig. 2.** For patterns that vary along five dimensions, these curves are the implied generalization gradients for three alternative stimulus representations: (1) component and pair-wise (doublet) cues, (2) component, doublet, and triplet-cues, and (3) component-cues only.

this network model can be shown to embody an exponential-similarity (generalization) gradient. This equivalence can be exhibited by computing how the number of overlapping active-input nodes (similarity) changes as a function of the number of overlapping component cues (distance). If a stimulus pattern is associated with some outcome, the configural-cue model will generalize this association to other stimulus patterns in proportion to the number of common input nodes they both activate. Figure 2 illustrates this relationship for stimulus patterns composed of five features. These patterns each activate 15 input nodes—5 component-cue nodes and 10 configural nodes. If two such patterns share only one feature (ABCDE, AWXYZ), they have 1 active node in common (A) and 14 nodes nonoverlapping. If they share two features (ABCDE, ABXYZ), they have 3 active nodes in common (A, B, AB) and 11 nodes nonoverlapping. If they share three features (ABCDE, ABCYZ), they have 6 active nodes in common (A, B, C, AB, AC, BC) and 9 nodes nonoverlapping . . . and so forth. The addition of triplet-cues further magnifies the upward concavity of the generalization gradient. Activating only the component cues results in a linear gradient.

## SHEPARD'S THEORY OF STIMULUS GENERALIZATION

Shepard (1987) describes a rational motivation for the exponential-decay generalization gradient that, he suggests, may account for its ubiquity in animal and human behavior. We show now that the emergence of this generalization gradient in the configural-cue model arises from assumptions isomorphic to a special case of Shepard's theory.

In a canonical-stimulus generalization experiment an individual is given a single reinforced trial with a training stimulus and is subsequently tested with a novel test stimulus. Shepard (1987) argues that the individual's behavior in this situation reflects an implicit estimate of the probability that the test stimulus, $S_T$, will have the same reinforcing consequences as the training stimulus, $S_0$. If the stimuli are represented within an appropriate psychological space (Shepard, 1958), sets of objects having common consequences are presumed to be similar to each other and to occupy well-behaved (e.g., compact and convex) regions of space called "consequential regions." Following a single training trial, the individual is presumed to have no knowledge of the size or location of the relevant consequential region, only that it contains $S_0$, the training item. To estimate the probability that the test item, $S_T$, lies in the same consequential region as $S_0$, the individual must consider all consequential regions that contain $S_0$, and evaluate the probability that each contains $S_T$. Shepard's principal result is that this analysis yields a close approximation to an exponential-decay gradient, regardless of the shapes of the regions or the distribution of the sizes of the regions.

### Consequential Regions and Configural-Cues

Shepard's approach can also be applied to understanding how the configural-cue model solves categorization problems. Although Shepard's theory of stimulus generalization was originally developed for continuous dimensions (such as tone frequency), it is equally applicable to stimuli composed of separable, discrete-valued stimuli (Shepard, 1989; see also Russell, 1986, 1988, for related statistical derivations of similarity among stimuli with discrete features). For example, consider the 8 possible patterns that can be composed from 3 separable, binary-valued feature dimensions. Figure 3A is a geometric representation of the psychological space containing these stimuli. If pattern [0,0,0] in the lower left corner is the memory item, $S_0$, then the three panels of Figure 3B show the faces, edges, and corners, respectively, which include $S_0$. These 7 regions correspond exactly to the 7 input nodes that will be activated in the configural-cue model by the presentation of $S_0$ (assuming the encoding of component, doublet, and triplet cues). Thus, the representation of stimuli in the configural-cue model is isomorphic to the "activation" of all possible consequential regions.

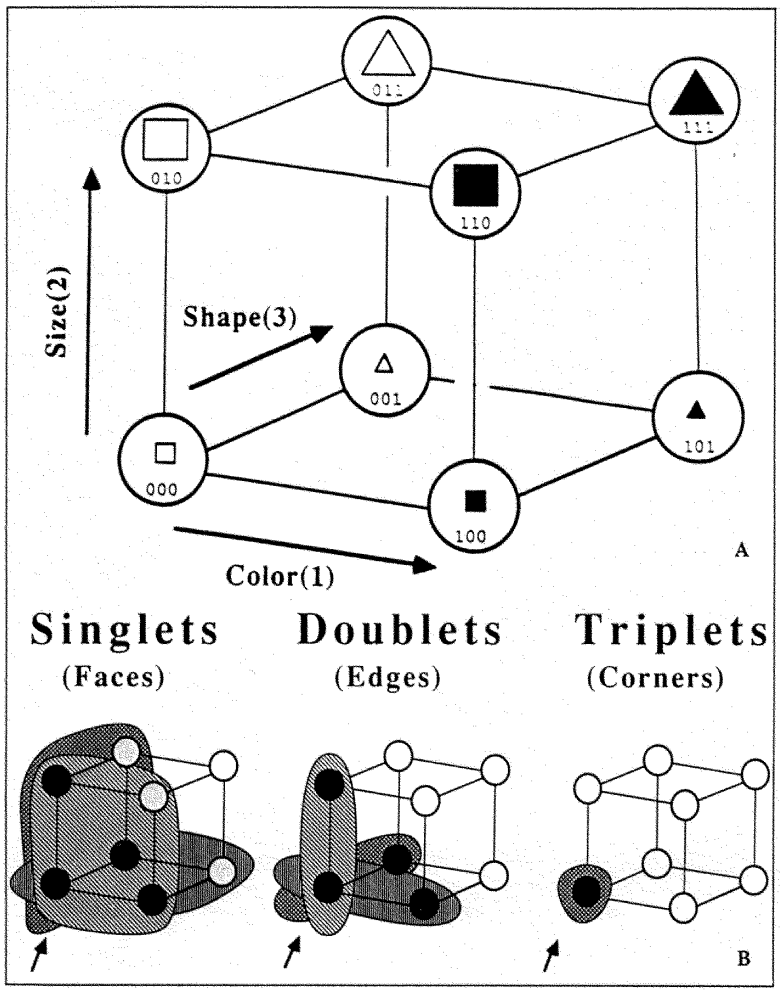


Fig. 3. (A) A geometric representation of the psychological state space for the 8 stimulus patterns that can be composed from 3 separable binary-valued feature dimensions. For illustrative purposes, we have identified the 3 geometric axes with the featural dimensions of color (dimension 1), size (dimension 2), and shape (dimension 3). (B) According to Shepard (1987), the faces, edges, and corners shown in these three panels are the consequential regions that will be "activated" by the association of stimulus pattern, [0,0,0], in the lower left corner with a consequential event.

## STIMULUS GENERALIZATION AND CLASSIFICATION: AN EXAMPLE

Data suitable to illustrate the implications of this nonlinear stimulus generalization gradient for classification learning are provided by Medin and Schwanenflugel (1981). They contrasted performance of groups of subjects learning pairs of classification tasks, one of which was *linearly separable* and one of which was not. The stimuli in their experiments were constructed so that the influences of exemplar similarity (when computed with an exponential-decay function) would favor the nonlinearly separable (NLS) over the linearly separable (LS) classification.

We begin with their Experiment 4, which used stimuli constructed from 3 binary-valued dimensions. Because these stimuli can be represented as the corners of a cube (Figure 3A), the exemplars of the two categories can be identified by differentially coloring the cubes' corners (Figure 4). This geometric representation makes the linear separability of the LS classification apparent; we can visualize a plane slicing the cube that partitions Category A (black corners) from Category B (white corners) in Figure 4A. No such plane exists for the nonseparable, NLS, task (Figure 4B).

To see why the nonlinear similarity rule of the exemplar-context model predicts that this LS task will be more difficult than the NLS task, we compare the within-category exemplar distances to the between-category exemplar distances for the two classifications. To calculate the average between-category distance for each task, we sum the Hamming (city block) distances from each black dot to each white dot in Figure 4, and take the average distance. Hamming distance is the minimal number of edges encountered en route from one corner (exemplar) of a cube to another corner (exemplar). This city-block metric is generally accepted as appropriate for psychological distances among discrete-valued separable stimuli (Nosofsky, 1984; Shepard, 1987). Calculation of the average within-category distances proceeds similarly. The LS and NLS classification tasks were constructed so that both tasks had identical average within-category distances (of 2), and identical between-category distances (of 5/3). Models that independently sum similarities along different dimensions (such as independent-cue, prototype models) predict that there should be no difference in difficulty of learning between the two tasks.

An important difference between these tasks is in the distribution of their between and within-category distances. The histogram in Figure 4A shows that the linearly separable task has many "close" (distance = 1) and some "far" (distance = 3) relations, whereas the nonseparable task (Figure 4B) has a broader distribution of "close," "medium," and "far" between-category distances. These disparate distributions have important implications for theories that exaggerate stimulus generalization between exemplar relations involving close distances: highly similar pairs of exemplars within a category should facil-

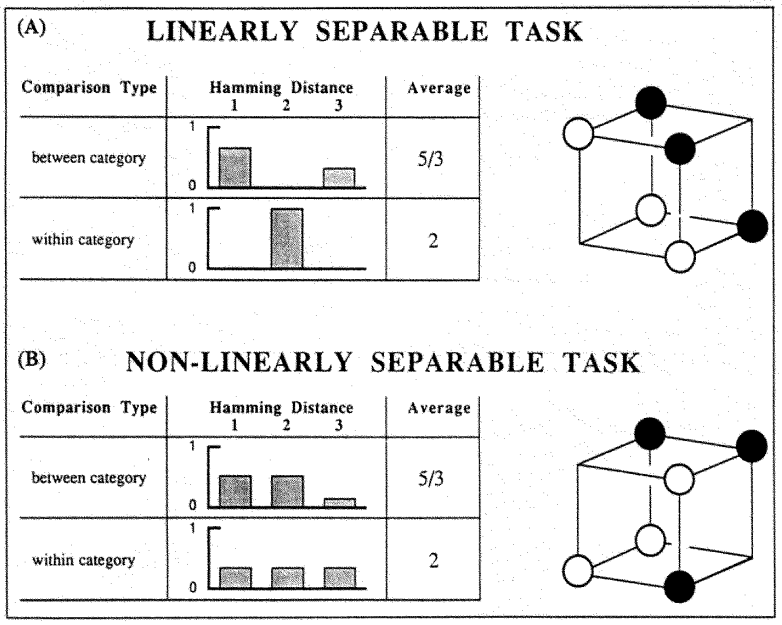


Fig. 4. A geometric representation of Medin and Schwanenflugel's (1981) Experiment 4. The histograms on the left show the distribution of between-category and within-category exemplar Hamming distances for each task. The cubes on the right show a geometric representation of each task; the black corners represent exemplars of category A and the white corners represent exemplars of category B. Panels A and B represent the linearly separable (LS) and nonlinearly separable (NLS) tasks, respectively.

itate classification learning, whereas highly similar exemplars belonging to different categories should retard learning. The greater proportion of these close between-category distances in the LS classification increases the confusion between A and B patterns; thus, the exemplar-context theory predicts that the LS classification should be more difficult to learn than the NLS classification. A similar analysis of within-category distances indicates the presence of fewer close distances in the linearly separable compared to the nonlinearly separable task; again implying that the linearly separable task should be more difficult.

It was convenient to use the three-dimensional classifications of Medin and Schwanenflugel (1981, Experiment 4) to convey through Figure 4 an intuitive understanding of the structure of the LS and NLS tasks, and the rationale underlying their design. However, Medin and Schwanenflugel reported more reliable and complete results with a four-dimensional task that embodied the same controls for linear separability and interexemplar similarities, and the data obtained from this four-dimensional stimulus structure (their Experiment 3) will be compared to the predictions of different classification theories. Table 1 schematizes the two groups of 6 stimulus patterns that college students learned to classify as members of category A or category B. The two values of each of the 4 binary-valued stimulus dimensions are denoted 1 and 0. To recognize the linear separability of the top (LS) classification, note that the number of 1s in dimensions 1, 3, and 4 is 2 for any category A stimulus, but is less than 2 for any category B stimulus.

Medin and Schwanenflugel (1981) compared the average learning curves of subjects trained on these classifications. Figure 5A shows that the results were as predicted by the context model: subjects found the linearly separable LS task to be harder to learn than the nonlinearly separable NLS task.

Because the configural-cue network implicitly embodies the same sensitivity to interexemplar similarities as the context model, we expect that it also should predict that the LS task will be more difficult than the NLS task. To test this prediction, we trained a configural-cue network (with component and doublet-cues) on the stimulus structure in Table 1. As shown in Figure 5B, this model correctly predicts that subjects will find the LS task more difficult than the NLS task. Although the model has 1 free parameter, a learning rate, the ordinal prediction that the LS task will be more difficult than the NLS task is parameter-free. The

**Table 1.** *Stimulus Design from Medin and Schwanenflugel (1978) Experiment 3*

| Linearly Separable (LS) Classification | | | |
|---|---|---|---|
| Category A | | Category B | |
| Exemplar | Dimension 1234 | Exemplar | Dimension 1234 |
| A1 | 0111 | B1 | 1000 |
| A2 | 1110 | B2 | 0001 |
| A3 | 1001 | B3 | 0110 |
| **Nonlinearly Separable (NLS) Classification** | | | |
| Category A | | Category B | |
| Exemplar | Dimension 1234 | Exemplar | Dimension 1234 |
| A1 | 1100 | B1 | 0000 |
| A2 | 0011 | B2 | 0101 |
| A3 | 1111 | B3 | 1010 |

simulation shown in Figure 5B used a learning rate of .013, chosen to minimize the squared deviation between the predicted and observed learning curves (RMSE = .037). The inclusion of additional higher-order terms (e.g., triplets) does not change the relative difficulty of the two tasks and results in no quantitative advantage in fitting the learning curve. Although the exemplar-context model makes the same ordinal prediction regarding the relative difficulty of the two tasks, it is not a learning model and, hence, cannot be used in its current form to generate learning curves that could be compared with those of the network model.

### Comparing Alternative Network Models

Several alternative network models can also be tested against this same data set. Clearly the original component-cue network model of Gluck and Bower (1988a) is not a viable model for these data; that model can never fully learn the nonlinearly separable NLS task. Estes, Campbell, Hotsopoulos, and Hurwitz (1989) proposed a different method for expanding the network's stimulus coding to enable it to solve nonlinearly separable classifications. Their "feature-pattern" model extends the component-cue representation by adding additional nodes representing the presence or absence of entire patterns. While capable of solving many complex discriminations, this model embodies the same linear generalization gradient as the component-cue model. In the feature-pattern model, the number of common input nodes activated by a training pattern and a test pattern in a stimulus generalization task equals the number of overlapping component cues only. It is not surprising, therefore, that the feature-pattern model incorrectly predicts that subjects should find Medin and Schwanenflugel's (1981) LS task easier to learn than the NLS task.

Another strategy for extending network models to nonlinearly separable classifications is to include an additional layer of "hidden units" between the input and output layer. Due to the large number of parameters and structural assumptions required to specify a multilayer network, it is not feasible to evaluate this class of models exhaustively.

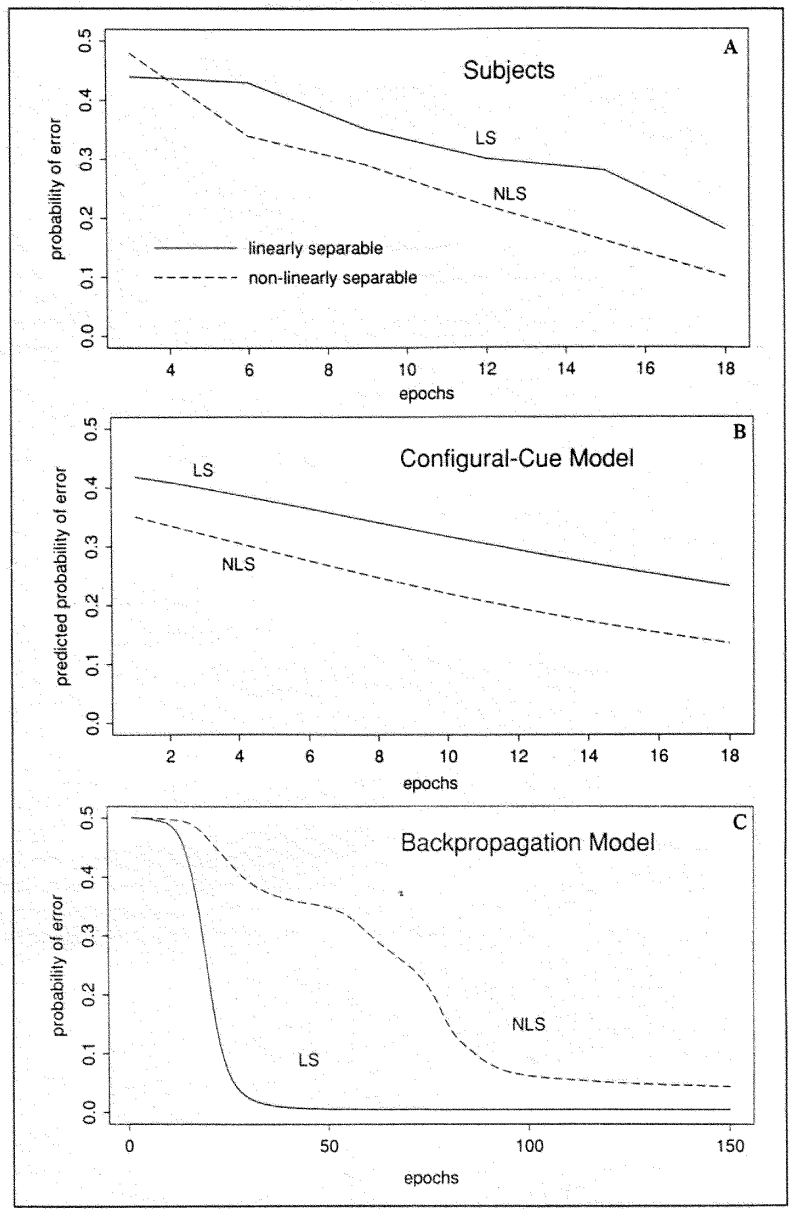


Fig. 5. The relative difficulty of the two classification tasks from Medin and Schwanenflugel (1981), Experiment 3. Output activations are mapped to a probability of a correct classification using a *ratio response rule* in which the rare occurrences of negative output activations are converted to 0. LS: linearly separable classification; NLS: nonlinearly separable classification. (A) The data on percentage errors, showing that the LS task is more difficult (slower to learn) than the NLS task; adapted from Medin and Schwanenflugel (1981). (B) The correct prediction of the "pair-wise" configural-cue model showing that the LS category is more difficult (slower to learn). (C) The incorrect prediction of a multilayer network with 2 hidden units. The same ordering of relative difficulty was also found with multilayer networks with 4, 8, and 16 hidden units, independent of the initial values of the weights or the learning-rate parameters.

Most current applications of multilayer networks, however, adopt stimulus representation assumptions similar to the component-cue model within a network whose hidden units are fully connected to both input and output layers. We adopted these same assumptions to evaluate a "base-line" multilayer network that we trained on the LS and NLS tasks of Table 1 using the "backpropagation" training procedure (Parker, 1986; Rumelhart, Hinton, & Williams, 1986; Werbos, 1974). As shown in Figure 5C, this model incorrectly predicts that subjects should find the LS task easier than the NLS task.

## GENERAL DISCUSSION

Shepard's (1987) theory of stimulus generalization applies only to the highly idealized experiment in which a single learning trial is followed immediately by a generalization test. The configural-cue network model can be viewed as an extension of Shepard's theory to discrimination and classification learning using the principles of associative learning from Rescorla and Wagner's (1972) model of classical conditioning. The successes of the configural-cue model in accounting for both animal and human learning can therefore be construed as independent and converging evidence for Shepard's theory.

This connection between theories of associative learning and theories of stimulus generalization suggests several new theoretical directions that might extend the range of phenomena deduceable from either theory alone. We briefly note three such possibilities here. First, Shepard (1987) has also shown that the implications of his theory are largely unaffected by the distribution of the sizes of the consequential regions. We conjecture that this result might be related to our observation that the predictions of the configural-cue model are largely unaffected by the addition of configural-cues more complex than pair-wise combinations or by most variations in the individual learning rates assigned to the configural-cues.

A second possible new direction is motivated by a serious limitation of the configural-cue model. In its current form it is applicable only to stimuli composed of separable discrete-valued features. Shepard's theory provides a broader theoretical framework within which we might identify stimulus representations described by continuous and integral feature dimensions.

Finally, a third possible research direction is suggested by the success of the configural-cue model compared to the base-line multilayer networks. In spite of their complexity, multilayer networks have the attractive property that they can dynamically reconfigure a small set of hidden units, thereby avoiding the configural-cue model's problematic assumption that input nodes exist, a priori, for all conjunctive-cue combinations. The evidently critical role of Shepard's stimulus generalization principles in the configural-cue model's ability to account for learning behaviors may point us toward the development of a categorization model that embodies these same generalization principles within a multilayer network.

**Acknowledgments**—For their thoughtful comments and advice on this work I am indebted to Gordon Bower, W.K. Estes, Michael Hee, Doug Medin, Robert Nosofsky, David Rumelhart, and Roger Shepard. This research was supported by ONR Grant #N00014-83K-0238 and by a grant from the Sloan Foundation.

## REFERENCES

Estes, W.K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W.K., Campbell, J.A., Hatsopoulos, N., & Hurwitz, J.B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 556–571.

Gluck, M.A., & Bower, G.H. (1988a). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 225–244.

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

Gluck, M.A., Bower, G.H., & Hee, M.R. (1989). A configural-cue network model of animal and human associative learning. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, Michigan, August 16–19, 1989* (pp. 323–332). Hillsdale, NJ: Erlbaum.

Gluck, M.A., Bower, G.H., & Hee, M.R. (1990). *Animal and human associative learning: A configural-cue network model.* Unpublished manuscript, Dept. of Psychology, Stanford University, Stanford, CA.

Gluck, M.A., Corter, J.H., & Bower, G.H. (1990). *Basic levels in the learning of category hierarchies: An adaptive network model.* Unpublished manuscript, Stanford University, Stanford, CA.

Kehoe, E.J., & Gormezano, I. (1980). Configuration and combination laws in conditioning with compound stimuli. *Psychological Bulletin, 87,* 351–378.

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

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear seperability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 104–114.

Pao, Y.H. (1989). *Adaptive pattern recognition and neural networks.* Reading, MA: Addison-Wesley.

Parker, D. (1986). A comparison of algorithms for neuron-like cells. In *Proceedings of the Neural Networks for Computing Conference* (pp. 327–332). Snowbird, UT:

Rescorla, R.A. (1972). "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology, 79,* 307–317.

Rescorla, R.A. (1973). Evidence for "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology, 85,* 331–338.

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

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D. Rumelhart, & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1: Foundations)* (pp. 318–362). Cambridge, MA: MIT Press.

Russell, S.J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 284–288). Philadelphia, PA: AAAI.

Russell, S.J. (1988). Analogy by similarity. In D.H. Helman (Ed.), *Analogical reasoning: Perspective of artificial intelligence, cognitive science, and philosophy* (pp. 251–269). Dordrecht, Holland: Klewer Academic Publishers.

Shepard, R.N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review, 65,* 242–256.

Shepard, R.N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Shepard, R.N. (1989). *A law of generalization and connectionist learning.* Talk presented at the Annual Meeting of the Cognitive Science Society, Ann Arbor, MI, August 17–18.

Smith, E., & Medin, D. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Wagner, A.R., & Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Applications of a theory. In R.A. Boakes, & S. Halliday (Eds.), *Inhibition and learning* (pp. 301–336). New York: Academic Press.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Doctoral dissertation (Economics), Harvard University, Cambridge, MA.

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, 4,* 96–194.

(RECEIVED 1/23/90; REVISION ACCEPTED 7/26/90)