To be Submitted to: *Psych. Review*

--- Manuscript in Preparation: Working Draft Only ----

A Configural-Cue Network Model of Associative Learning in Animals and Humans

Mark A. Gluck
Center for Molecular and Behavioral Neuroscience
Rutgers University

Gordon H. Bower Michael R. Hee
Department of Psychology
Stanford University

Running Head: "Configural-Cue Network Model"

February 12, 1992

Abstract

NOTE: This is from cog sci conf paper

Too cryptic and rapid -- GB

1

We test a configural-cue network model of human classification and recognition learning based on Rescorla & Wagner's (1972) model of classical conditioning. The model extends the stimulus representation assumptions from our earlier one-layer network model (Gluck & Bower, 1988b) to include pair-wise conjunctions of features as unique cues. Like the exemplar context model of Medin & Schaffer (1978), the representational assumptions of the configural-cue network model embody an implicit exponential decay relationship between stimulus similarity and and psychological (Hamming) distance, a relationship which has received substantial independent empirical and theoretical support (Shepard, 1957, 1987). In addition to results from animal learning, the model accounts for several aspects of complex human category learning, including the relationship between category similarity and linear separability in determining classification difficulty (Medin & Schwanenflugel, 1981), the relationship between classification and recognition memory for instances (Hayes-Roth & Hayes-Roth, 1977), and the impact of correlated attributes on classification (Medin, Altom, Edelson, & Freko, 1982).

Introduction

This paper presents and tests a new, simple model of human classification learning. Its primary emphasis is upon explaining the results of many laboratory experiments in which adults learn to classify a collection of stimulu~~a~~s patterns (or "cases") into two or more categories based on feedback about each pattern.

Classification or categorization is a fundamental cognitive act, and in some sense it underlies all learning, all transfer or generalization, and most induction of generalizations. People use multiple classifications of their environment in nearly every action they perform. Classification involves treating a person, object, place, or situation as equivalent for some purposes to one or more experienced before. ~~Because of~~ This judgment of equivalence, ~~we~~ permits subjects to ~~can~~ partially transfer to the new case some of the knowledge ~~we~~ they had gleaned from the earlier cases. For even the most elementary habituation or classical conditioning experiments, the human or lower animal may be said to be classifying the present stimulus as resembling one experienced earlier, and responding accordingly ~~to that judgment.~~

Because of its fundamental nature, classification -- its nature and learning -- has long been studied by psychologists under such headings as discrimination learning, pattern recognition, concept identification, and categorization. As more data have accumulated, considerable theorizing has grown up around this body of results. Despite the accumulated data and theoretical elaborations, no consensus has yet been reached about the best theory. Into this area, we offer our own theory, called the "configural cue" theory. In the following, we describe the basic assumptions of this theory, show its connections to traditional approaches, contrast it to popular alternative models of category learning, and summarize some evidence in favor of our theory. Our particular emphasis will be upon formulating the configural cue theory in an explicit manner enabling us to derive its numerical predictions for a large collection ~~vast arrangemen~~t of category learning experiments. The model can then be evaluated by its fit to numerical results from an array of critical experiments. Moreover, the alternative models may be contrasted in their goodness of fit to the critical results.

Last step: insert preview of paper here

THE NETWORK-LEARNING FRAMEWORK

The Pandemonium Approach

The model to be presented combines three separate approaches to discrimination learning or pattern classification. The first one we call the "Pandemonium" approach, deriving as it does from the familiar Pandemonium-model of pattern classification. That approach makes at least four assumptions: first, that objects or patterns to be classified can be characterized by a collection of stimulus features or attributes which they possess; second, that each feature provides varying degrees of "evidence" for classifying the pattern into the different classes of interest, and that through learning this evidential degree comes to be reflected in the strengths of associations between the feature and the alternative categories; third, that the total weight of evidence towards a given classification for a pattern is given by summing the evidence each of its features provides towards each category; and fourth, that the ranking of amount of evidence for each category is converted into a behavioral decision about which "category response" to give to the stimulus pattern.

While these assumptions were transparent in Selfridge's Pandemonium Model (e.g., Selfridge & Neisser, 1960), they were in fact present as well in earlier treatments of discrimination learning by Kenneth Spence (1936), Clark Hull (1952), and most others working in neobehaviorism. For example, in accounting for a chimpanzee's selection of a *large black square covering the left food well* over a *small white circle covering the right food well* in a Wisconson General Test Apparatus, Spence's theory would add up the strengths of "approach habits" towards the four stimulus features on each side, and then convert those two net habit strengths to the two choices into a probability of the animal selecting one pattern over the other. The theory operated similarly when applied to a successive discrimination task, in which the organism learns to make response R_j to stimulus pattern P_i; the associations to each response from each attribute of the pattern would be summed over attributes, and the model's response would be selected according to its net strength relative to other responses. This type of decision rule is familiar in the literature on linear-regression models of statistical decisions (see Dawes, 1979) or the lens model of multiple-cue probability learning (e.g., Hammond et al., 19@@). Our point in citing these examples is to note that this assumption, that decisions are made by comparing the summation of evidence towards different alternatives, is familiar and widely used in psychological theories.

A specific network that we used in earlier publications is shown in Figure 1; it is a

----

Insert Figure 1 about here

----

simple "one-layer" network, associating input elements to output units in one pass. The input layer includes a sensory unit or "node" for each stimulus feature; for example, with patterns of colored geometric forms, ~~there~~ we would ~~be~~ postulate one sensory node corresponding to each distinct color and one corresponding to each form used in the experiment. Presentation of a given stimulus pattern to the network is represented ~~by~~ as activating (turning on) the input nodes corresponding to the presented features of that pattern. Mathematically we will let a_i represent the amount of activation stemming from input node i when its corresponding stimulus attribute is presented. Intuitively, a_i ~~may be conceived of as~~ reflects the salience or intensity of that sensory feature relative to others. In the following, however, we will set all a_i to the same value (namely, 1) whenever the corresponding feature is present, and set a_i to 0 if that feature is not present. Later, we will complicate the nature of the sensory input; but the elementary model of Figure 1 serves as a useful starting point for exposition.

Each input node, i, is connected to each output (category response) node j with a weight or strength of degree w_{ij}; we think of this as the directed association from i to j, corresponding to the degree of evidence towards category j provided by feature i. It is these w's which will change to reflect the stimulus-response correlations in the training environment. Given an input pattern, the network computes the activation on each output node by summing the associative weights to that response node from each presented stimulus feature. Letting A_j denote the activation or summed strength of output node j, it is given by

$$A_j = \sum_i a_i \ w_{ij} \ . \qquad (1)$$

Recall that a_i is one if feature i is presented in the pattern on the current trial but is zero otherwise.

We will add one further assumption to our network model, and that regards the specific rule for selecting a "response" or classification. Letting A_1, A_2,......, A_n represent the total strength of associations from a pattern to category responses 1, 2, ...n, then the probability that the person selects response A_i out of a set of alternatives including it will be assumed to equal

$$P_i = \frac{A_i}{\sum_i A_j} . \qquad (2)$$

The summation in the denominator ranges over just the set of alternatives available for choice on a given trial. Equation 2 is Luce's Axiom which has many desirable features (Luce, 1959). For example, one feature is that the ratio of P_i to P_j is predicted to be independent of the set of alternatives in which those two are included -- a prediction that is frequently upheld in choice experiments (see, e.g., Atkinson, Bower, & Crothers, 1965).

THE LMS LEARNING RULE

The second major assumption in our network model is that new learning is to be driven by the failure of prior expectations; that is, new expectations are adjusted to reduce discrepancies between old expectations and reality. This process of discrepancy-reduction is evoked whenever the organism expects some outcome to a degree that does not match the actual outcome. Although we believe that this expectation learning proceeds amongst all features of the stimulus patterns as well as from the stimuli to the classificatory responses, we will concentrate here on how the sensory units become associated to the output responses.

The model operates in a training environment consisting of a number of discrete trials during which a stimulus pattern is shown, the network calculates activations (or strengths of expectations) for the available classificatory responses and responds, and then a supervisor, critic, or trainer provides feedback regarding the classification (or "correct response") that should have been given on that trial. We suppose that the feedback signal informs the network of what the desired activations over the N response alternatives should have been on that trial. For the output node j, we let λ_j denote the training signal given on a specific trial. In general, the value of λ_j would vary with parameters of reinforcement, such as its specificity, magnitude and delay. However, when reward is not varied, as in all the experiments considered below, we will use the simplest case in which λ_j is set to some arbitrary value (namely, 1) for the correct category (response) node on a given trial and 0 otherwise. We will let $w_{ij}(n)$ denote the weight of the connection from cue i to category j at the beginning of trial n of some training series. These w_{ij}'s will change depending on the training schedule. The learning that follows a given feedback trial can now be stated as follows:

NO

BEST

$$w_{ij}(n+1) = w_{ij}(n) + a_i\beta\left[\lambda_j - \sum_i a_i w_{ij}\right]. \tag{3}$$

<u>If we add trial subscripts they need to be added through EQ-MG</u>

Equation 3 is the central learning axiom of the theory. The parameter β is the learning rate and is to be estimated from fitting experimental data; typical values are .01 to .03.

Several features of Equation 3 should be noted. First, it really synopsizes three underlying relationships, namely,

$$w_{ij}\left[n+1\right] = \begin{cases} w_{ij}(n) + \beta\left[1 - \sum_i a_i w_{ij}\right] & \text{(3a) if feature i was presented and response j was correct} \\ w_{ij}(n) - \beta\left[\sum_i a_i w_{ij}\right] & \text{(3b) if feature i was presented and response j was not correct} \\ w_{ij}(n) & \text{(3c) if feature i was not presented or there was no feedback} \end{cases}$$

BIG

The first line (3a) says that reinforcement of response R_j increases the association of the presented stimulus features to that response or category. The second line (3b) says that reinforcement of some other competing response lowers the association of feature i to R_j. The third line (3c) says that associations do not change either if the corresponding stimulus is not presented on this trial or if ~~there's~~ no training signal or reinforcing feedback is provided on a given trial.

Second, if the experiment consists solely of repeated presentations of a single pattern that is consistently reinforced with the same correct response, then Equation 3a implies the standard exponential learning curve of Hull's (1943) theory or that of Estes (1950).

Third, the changes in weights specified by Equations 3a or 3b are the same for all presented stimulus features, regardless of their current strength.

Fourth, the weight changes are proportional to the discrepancy (delta, difference) between the current output activations and the desired outputs, the λ_js. The result of such weight changes is to reduce the discrepancy were this same pattern to be immediately presented again.[2]

Fifth, Equation 3 is variously called the delta rule (for reducing differences), the Widrow-Hoff rule (after its originators in linear systems theory, Widrow and Hoff (1960)), and the Least-Mean-Squares (LMS) rule. The LMS appellation stems from the fact that Equation 3 provides an iterative method to calculate the weight coefficients so as to minimize asymptotically the expected squared errors of actual versus desired outputs when averaged over the patterns in the training set. The weights are similar to regression coefficients in a linear regression model that aims to predict several criterion variables (the λ_j) from a weighted sum of independent variables, namely, the stimulus features, represented by the a_i's. Like regression weights, the w_{ij} of Equation 3 will eventually reflect the partial correlations between input features and the category outcomes, once the coefficients are corrected for inter-correlations among the input features themselves.

Sixth, as noted by Sutton and Barto (198x), Equation 3 has the same form as the rule of Pavlovian conditioning proposed by Rescorla and Wagner (1972). The notations in our Equation 3 differ slightly from those of Rescorla and Wagner (e.g., they used V's rather than w's to denote associative strength), but the sense of the rule is the same. Rescorla and Wagner proposed the rule as a description of a number of striking phenomena observed in classical conditioning (see below), and the rule has a remarkably broad range of predictions that have been confirmed in the conditioning laboratory. The next section reviews several of these implications.

Conditioning Evidence for the LMS Learning Rule

The strongest evidence for the LMS learning rule comes from the phenomena of *blocking* and *overshadowing* in studies of classical conditioning. In Kamin's (1969) initial study of blocking, rats first associated a neutral tone with a painful footshock, so that the tone came to produce conditioned suppression of bar pressing for food. The rats then experienced trials with the stimulus compound, tone-*plus*-light, paired with the same shock. A later test to the single stimuli showed a continued high level of fear evoked by the tone alone, but virtually no conditioning of fear to the light alone. Apparently, the prior conditioning of the tone to shock had blocked the learning of light to shock when the tone-*plus*-light compound was paired with shock. In contrast, other animals who had only trials of tone + light paired with shock (without the earlier tone-shock pairings) showed appreciable fear when tested with the tone alone or the light alone.

Such blocking is predicted by the LMS rule. Notice that during the initial phase of the experiment the tone alone is fully conditioned to the shock, so that by Equation 3a the associative weight of tone to shock, w_{TS} moves to near whatever strength, λ, is supported by the shock level. Later when the same shock is paired with the tone-plus-light compound, with the initial light-shock association starting at zero, Equation 3a tells us that the change in the light-shock association will be zero. In this manner, the blocking effect is implied by the LMS rule.

The blocking phenomenon and its causes have been described in several different ways. All descriptions include the idea that simple contiguous conjunction of a cue and an unconditional stimulus is not sufficient to produce associative learning. Rather, to become effective, a cue must impart nonredundant, predictive information about the upcoming unconditioned stimulus beyond that provided by accompanying cues or contexts. One interpretation of blocking is attentional, that weak cues receive less attention, hence receive less conditioning because attention is taken up by co-present strong signals for the outcome. An alternative interpretation refers to a weakened reinforcing power of the US during blocking, namely, that an expected US does not have the same potency for promoting new associations as does an unexpected US. For example, Wagner (19@@) uses the gradual habituation of the unconditioned response of animals to a constant shock in a constant setting as an indicator of this loss of potency of an expected US.

Overshadowing is a similar phenomenon observed when conditioning trials with a single stimulus, say a tone, occurs in random alternation with conditioning trials involving a compound stimulus including the former, say that same tone, plus a light. According to Equation 3a, the result of the single tone-shock pairings is to drive w_{TS} towards the asymptote λ; consequently, the discrepancy on tone-plus-light trials, which is $\lambda - (w_{TS} + w_{LS})$, will be smaller the larger w_{TS} becomes. As a consequence, the change in the light-shock association, w_{LS}, will be much less than it would were only the compound trials to occur without the single tone-shock trials.

Inhibitory conditioning is also implied by the LMS rule. In this case, the organism learns to expect no unconditional stimulus in a prevailing context where it has often occurred. To reflect inhibitory conditioning, the weights are allowed to take on negative values as indicated by Equation 3. Such negative values arise in paradigms known to produce conditioned inhibitors. The existence of conditioned inhibitors implies that one should be able to "super condition" a new cue placed into compound with an inhibitor but with the compound now paired with the US. If cue B is an inhibitor of shock expectancy, then the change in conditioning of a novel cue, C, when the compound BC is paired with the US is expected to be very large. This superiority dubbed "superconditioning," has been confirmed several times (Rescorla@,@@,).

While the Rescorla-Wagner theory is supported by a more extensive range of conditioning phenomena, we move on to review of studies relevant to the LMS rule operating in human earning, since that is the domain classification learning.

Blocking and Overshadowing in Human Learning

Some evidence suggests that adult human subjects show effects similar to blocking and overshadowing. An early demonstration of blocking was reported by Bower and Trabasso (1964) studying college-student subjects learning in a concept identification task. Subjects were presented with geometric figures varying in their color, size, shape, and embroidery. During an initial learning phase, shape was the relevant dimension with circles placed in Class A and triangles in Class B, with all other variable dimensions irrelevant. After initial learning, an overtraining phase of 32 trials followed during which color was made redundant and relevant along with shape, e.g., red circles were As and green triangles were Bs, and neither red triangles nor green circles were presented. Thereafter, subjects were assessed for how they would classify various incomplete patterns, such as a red figure whose shape was not specified or an outlined circle without color. Such tests revealed that the earlier learning of the shape-to-category associations blocked the later learning of associations to the color cues that had been made relevant and redundant during overtraining. That is, having just classified 16 red circles as A's and 16 green triangles as

No: it was
dot above/below
central figure

B's, subjects were unable to guess above chance how to classify red or green figures whose shape was unknown.

A more recent demonstration of *blocking* has been reported by Chapman (1991). College-student subjects were presented with a series of trials in which the stimulus patterns were framed as patients exhibiting 0 to 6 medical symptoms. For each patient, subjects estimated the probability that the patient had a specific fictitious disease, and then were told whether or not the patient had the disease. Periodically subjects were asked to estimate the probability that a patient with a specific symptom (with no information about presence or absence of the others) had the disease. In Phase 1 of the experiment, one symptom was a perfectly valid predictor of the disease while another was a valid predictor of absence of the critical disease. These were called the P (positive) and N (negative) cues, respectively, since they were exptected to acquire positive and negative (inhibitory) associations to the disease. ~~respectively.~~ Other cues, L, M, K, etc. were presented but had no correlation with presence vs. absence of the disease. In the second phase of the experiment, three compounds of two symptoms were presented, all of which were completely valid predictors of the disease: PA, NB, and MC. At the end of the second phase, subjects rated the single cues P, A, N, B, M, and C ~~were evaluated~~ for their individual validities for predicting the disease. Equation 3 expects that the strong association established earlier for cue P will block the learning of the cue A-to-disease association compared to the learning of cue C in the context of the neutral MC compound. On the other hand, the earlier inhibitory conditioning (or expectation of no-disease) of cue N will enhance the learning of the cue B-to-disease association, relative to the association of cue C -- a form of "super conditioning." Exactly these results were obtained by Chapman: subjects' final estimates of the probability of the disease given the single symptoms fall in the order B (highest), then C, then A (lowest). The essentials of these results were replicated in further experiments by Chapman.

Several other human learning experiments also demonstrate *overshadowing*. An experiment by Shanks (1991) is illustrative. In one experiment, subjects learned to associate a compound of medical symptoms, AB, with Disease #1 and another compound, CD, with a Disease #2. In random alternation with these compound trials, subjects also received trials with single-symptoms, including symptom B alone predicting absence of all disease, and symptom D alone also predicting Disease #2. Notice that by Equation 3, these arrangements imply that cue A will be more strongly associated with Disease #1 (because cue B has no such associative strength), whereas cue C will be weakly associated to Disease #2 (because it is always in compound with cue D which also predicts that disease). On tests with individual symptoms given at the end of training, Shanks confirmed this prediction: subjects rated the probability of Disease 1 given cue A nearly twice as high as the probability of Disease 2 given cue C. Chapman and Robbins (1990) have reported similar findings.

A further implication of Equation 3 is that cues will come to be ordered in their asymptotic strength according to their validity for predicting an outcome *relative* to that of other co-occurring cues. Experiment 3 of Shanks (1991) may be used to illustrate this principle. The experiment, fashioned after an earlier one by Wagner, Logan, Haberlandt, ~~&~~ and Price (1968), compared the learning of elementary cues (medical symptoms) that were presented in one of four compounds. Symptom compounds AB and AC were presented equally often with the first always paired with Disease #1, and the second always paired with absence of the disease; compounds DE and DF were both paired equally often with presence vs. absence of Disease #2 ~~or with absence of disease~~. After training, subjects rated the degree of association of individual symptoms to the various diseases. Interest centers on the comparison of the associative strengths of cue A versus cue D to their respective diseases. Note that the validity of cue A for predicting Disease #1 is .50; that is, half the time cue A is present, Disease #1 occurs. Cue D similarly is partially valid; the probability of Disease #2 is .50 when cue D is present and is zero otherwise. However, the cues differ in that cue A always appears in the presence of a cue~~s~~ that ~~are~~ is more predictive of the outcome than is cue A, whereas cue D only appears with cues E and F that are no more predictive of Disease #2 than is D. In fact, D is advantaged because it occurred twice as often as E or F considered alone. Equation 3 implies that the strength of the cue A-to-Disease 1 association will be driven to a lower level than will the strength of the cue D-to-Disease 2 association. Shanks' findings confirmed this prediction, that cue D was stronger than cue A for their respective diseases. The important point here is that the conditioning of cue A was weaker than that of cue D because A always occurred in the presence of cues which were more predictive of the outcomes than it was. The result illustrates cue competition during learning: when cues appear together in compounds, more valid cues prevent less valid ones from acquiring much associative strength.

Cues Compete; Effects Do Not

An important caveat regarding the LMS rule is that the competition is explicitly restricted to cues that are simultensouly present and involved in predicting some outcome. This is the natural temporal ordering for cause-effect relations. In contrast, there is no competition involved if the stimuli are being used to predict several simultaneous outcomes that are not mutually exclusive. For example, if from ~~a person's~~ people's occupational and economic status we learned to predict ~~his or her housing situation~~ their food preferences that would not block the later learning to also predict their political affiliations. Experiments by Baker & X (198x) and Waldemann and Holyoak (in press) have shown that a few (usually two) multiple outcomes can be predicted from single causes without competition amongst the outcomes, whereas multiple predictor~~s~~ cues ("causes") do compete according to the LMS rule for association to a single effect. The reason for this asymmetry can be seen in the network of Figure 1: the LMS rule adjusts weights of links coming *into* a given outcome node; but the model allows several independent outcomes (e.g., a person's food preferences and political affiliation) to have their weights adjusted simultaneously without competition. ?

Causality Judgments and the LMS Rule

Recall that the concept of association was used by the philosopher, David Hume (1964), to explicate people's beliefs about causality (see Wasserman, 198@). The temporal conjunction of a cause with its effect was presumed to form a mental association between their ideas; a stronger sense of causal power was presumably produced by more frequent and more highly correlated causal conjunctions, which would be reflected in the strength of the underlying association.

Modern investigators have used several paradigms to study people's causal judgments. In one paradigm, subjects observe a lengthy series of antecedent-consequent pairs of events and then estimate the degree of causal relatedness of the events. An example is a study by Wasserman (1991) in which subjects tried to judge which foods cause an allergic reaction in a given patient. Subjects receive as data many observations of days when the patient ate certain combinations of foods and either did or did not have an allergic reaction as a consequence on that day. In general, the greater the correlation between presence/absence of a food eaten and presence/absence of the allergic reaction, the higher were subjects' judgments of the degree of causal relatedness of the food and the allergic reaction.

In a second paradigm, subjects make an operant response or not in each time interval and some outcome occurs or not at the end of the interval; after many trials, subjects judge the degree to which their response causes (or "controls") some outcome. For example, in a video game battlefield, subjects may control the firing of cannon shells from their unreliable artillery, aiming at an enemy tank crossing a randomly-laid mine field. ~~The~~ A given tank may or may not explode, and it may be caused by either the cannon shot or the land mines. Subjects judge the extent to which their cannon shots are causing the tank to explode.

Several reviews of the causal judgment literature have been written (Alloy & Tabachnick, 1984; Shanks & Dickinson, 1987; Wasserman, 1990). The major finding of many of these experiments is best explained using a contingency table (Table 1). In Table

----
Insert Table 1 about here

1, a, b, c, d denote the relative frequencies of the joint events in each cell; these probabilities sum to one. The consistent finding is that the impression of causal relatedness of event A to outcome B varies directly (proportionately) with the difference in the conditional probabilities of outcome B given A versus given not A, which is

$$\Delta p = \frac{a}{a+b} - \frac{c}{c+d}.$$

For example, when outcome B always follows event A but never arises otherwise, then $\Delta p = 1$ and the events convey a strong impression that A causes B.

For present purposes, the interesting aspect of this result is that it corresponds exactly to what the LMS learning rule expects when the four types of trials in Table 1 are mixed with the indicated probabilities. This relationship was first derived by Chapman and Robbins (1989). The average change per trial in associative strengths to cues A and X may be derived for a hypothetical experiment in which the four cue-outcome possibilities occur with probabilities a,b,c, and d. The averaged results ~~is~~ are :

$$\Delta w_x = \beta \left[a+c - (a+b)w_A - w_x \right]$$

$$\Delta w_A = \beta \left[a - (a+b)(w_A + w_x) \right].$$

At asymptote, the average change scores are zero. Solving these two equations, we find the asymptotic values of associative strengths to be

$$w_x = \frac{c}{c+d}$$

and

$$w_A = \frac{a}{a+b} - \frac{c}{c+d}.$$

Importantly, the implied associative strength of cue A for outcome B is Δp, which was as noted above. We may presume that empirical judgments of causal relatedness are linearly related to such associative strengths.

Other writers have noted the relation of the causality judgment literature to attribution theory as used in person perception studies (Alloy & Tabachnick, 1984; Kelley, 1967; Jones & Davis, 1965). Attribution theory describes how a social perceiver infers causes of a person's behavior from an implicit analysis of variance which notes how behaviors vary across situations and/or persons. Thus, if Richard drinks excessively in all situations, we attribute his behavior to an internal factor of alchohol addiction. But if he drinks excessively only at fraternity parties, we attribute such actions externally to the party situation. Similarly, when Richard donates blood during the Blood Drive, we attribute the cause more to his personal beliefs or motivations if only he contributes than if his entire fraternity has a tradition of frequently contributing. In each case, the social perceiver uses a "discounting" principle, whereby the inference from the target behavior to an underlying trait (cause) is blocked if the person's action can be explained by situational determinants or by baserates ("everyone does it").

Discounting also arises in inferences regarding causes of achievement. A successful or failed response may have come about due to internal factors such as the actor's ability and motivation (or effort), or been caused by external factors such as an easy or difficult task and good or bad luck (see, e.g., Weiner, 1986). The more an achievement is attributed to one factor, the less causal power the other factor is perceived to have.

We mention these phenomena because attributional discounting is a form of overshadowing: the judged potency of a possible cause for a given outcome is scaled back due to competition from a stronger causal factor present in the situation. Importantly, Equation 3 provides a general rule describing the extent of discounting found in attributional studies of social perception.

Stimulus Representation

The third major assumption in applying any network model concerns how to represent the stimuli or stimulus patterns of the experiment. In general, network models assume a collection of sensory units that are turned on by a stimulus pattern, and these send activation into the associative network. But there are a variety of identifications of what the "sensory units" or features are. In ~~many~~ some models such as TODAM (Murdock, 1982), CHARM (Metcalf-Eich, 1982), MINERVA (Hintzman, 1986), and similar ones, the presentation of, say, a colored geometric figure or a familiar word to an adult human would correspond to turning on an array of features, but these features are left unspecified and totally abstract. For certain purposes this vagueness suffices (see, e.g., the multicomponent model of Bower, 1967). But these models then suffer indeterminances when one tries to apply them to experiments involving novel recombinations, compounds, or concatenations of the experimentally-defined stimuli.

An alternative is to identify each single observable feature of an experimentally-defined stimulus with a corresponding input unit in the model. For example, in a classification experiment, presentation of a *large red triangle* would correspond in the model to turning on three sensory units representing the features *large, red,* and *triangle*. The complement, a *small blue square*, would correspond to turning on three *different* sensory units representing *small, blue,* and *square*. We used these identifications in several earlier papers, applying the model to our experiments on adults learning of probabilistic classifications (Gluck & Bower, 1986, 1988a, 1988b). For certain kinds of classifications, those simple representations work effectively and the model fits such data reasonably well (see also Estes et al., 199@; MacMillan, 1987; Nosofsky, 199@). Such representations do best in fitting experiments in which the category to which a stimulus pattern belongs is determined by the summation of evidence provided by each of the stimulus features considered singly and independently, and classification does not depend on the particular configuration of features present.

not here- earlier

Standard Prototype Results

It is appropriate to note that even the simple network model can explain most of the results originally offered in favor of prototype models (e.g., Franks & Bransford, 1971; Hintzman, 1986; Reed, 1972). In these experiments, a central prototypic member of each category is arbitrarily defined by a specific configuration of features; other instances of the category are generated by changing one or more of these characteristic features. Subjects usually are trained with several variants of each prototype category, and then are tested for transfer of the classifications to novel instances at varying "distances" (of mismatching features) from the prototype.

To outline the derivation while simplifying, by assuming equal treatment for all prototypic ~~and~~ features and also for nonprototypic features, we may then describe any test pattern in terms of the proportion, q, of prototypic features. Applying Equation 3 during training, Each prototypic feature will converge ~~during training~~ to asymptotic weights towards the target and null categories of w_{pT} and w_{po}, with $w_{pT} > w_{po}$; similarly, each nonprototypic feature will converge to weights of w_{nT} and wno, with $w_{nT} < wwubno$. A test pattern containing a proportion q of prototypic features will thus give rise to output activations given by

$$Target\ category = A_T = qw_{pt} + (1-q)w_{nT}$$

and

$$Null\ category = A_o = qw_{po} + (1-q)w_{no}.$$

The overall acceptability of a given test pattern is related to the ratio $A_{T/(A_T} + A_o)$. For the weight inequalities above, this ratio increases with q, the proportion of prototypic features in the test pattern. ~~Thus,~~ This implies that subjects will be more likely to classify a test pattern in the target category the greater the proportion of features it shares with the central prototype. This result, dubbed the "distance from prototype" effect, is often observed. An implication is that, other things being equal, subjects will best classify the prototype itself even if they have not previously seen exactly that combination of features. Similar to the distance effect on classification, subjects will rate the "typicality" of particular exemplars of the category according to the ratio of A_T to A_o.

Another result of the prototype-learning experiments is that the degree of transfer of a classification to new variants of a prototype increases the greater the number and range of variants experienced during training (e.g., Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Homa & Cultice, 1984). This result is understandable simply in terms of sampling probabilities: the greater the number and variety of training exemplars of a category, the more probable it is that a new test variant will be somewhat close (in terms of overlapping features) to one or more of the training exemplars, and so benefit from such similarity (of associated, overlapping features) in the classification.

A set of results that were problematic for the prototype theories were those suggesting that people's classifying is greatly influenced by the relative frequencies with which specific instances of a prototype category are presented (Smith & Medin, 1981).

The problem can be illustrated by an extreme case: suppose that only two very different variants of the experimenter's prototype are presented. It is typically observed that subjects learn these specific instances and respond to new cases not from some abstract average but rather according to their similarity to one or another of the training instances. ~~The difficulty is that~~ Clearly, subjects ~~must~~ have to be exposed to sufficiently many variants before the it is reasonable to define or abstract ~~notion of~~ a shared prototype ~~can be defined or abstracted~~ in a sensible manner. On the other hand, exemplar storage models (such as the context model of Medin~~e~~ & Schaeffer, 197x) track rather well the behavior of~~g~~ subjects exposed to varying frequencies of small numbers of instances. The extended network model discussed below also performs reasonably well in these circumstances which historically have caused difficulties for prototype-averaging theories of classification.

Linear Separability

The tact taken above was to identify each experimental stimulus or cue with a single input node in the model. However, a fundamental difficulty with such a model is that it can only learn perfectly those classifications that are "linearly separable" in terms of the single features. Essentially these are discriminations for which summation of evidence from single features considered independently provides all the information needed to respond appropriately to a stimulus pattern comprised of those features. Mathematically, linearly separable categories are those whose patterns can be perfectly segregated by comparing a weighted linear combination of each pattern's features to a threshold.

The concept of linear separability can be illustrated visually for two classes of patterns formed from two binary valued stimulus dimensions. Figure 2A represents the four possible patterns by the four corners of a square. Any two category tasks can be represented ~~described~~ by coloring the exemplars of one category black, and the exemplars of the other category white. In this case the black and white patterns are said to be linearly separable if a straight line can be drawn through the plane which separates the two sets of paterns. Figure 3A illustrates a linearly separable task (the OR task), while Figure 0B illustrates the simplest non-linearly separable classification (the exclusive-OR). With three stimulus dimensions, the eight patterns containing three binary valued component cues can be represented as the

----

Insert Figure 2 about here

Insert Figure 3 about here

----

corners on a cube (Figure 2B). Here, linear separability corresponds to the ability to draw a plane through the cube which separates the black from the white corners. Figures 3C and 3D illustrate ~~this representation for~~ a 3-dimensional separable task and a non-separable task, respectively.

The Configural-Cue Assumption

As noted, the simple network of Figure 1 using single cues as input nodes cannot learn categories that are not linearly separable in terms of those features. However, the problem can be solved by a simple reformulation, namely, by expanding the theoretical feature set to include various *conjunctions* of the elementary stimulus components, such as pairs (doublets) and triples (triplets) of presented features. Given the presentation of a stimulus pattern consisting of elementary features BCD, we assume that this ~~is reflected in~~ event gives rise to activation of input nodes corresponding to the single elements B, C, and D, and the pair-wise conjuncts BC, BD, and CD. ~~We call these~~ These conjuncts will be called "configural cues." By adding cue-configurations, the expanded model converts a classification which is nonlinear in the original singleton features into a classification that is linearly separable in terms of the expanded feature space. A network model containing only single and pairwise configural cues cannot perfectly learn non-linearly separable classifications that depend on third order (or higher) cue interactions. However, *n* th order cue configurations can be added to allow such models to perfectly discriminate such classification tasks. In practice, we have found that second order coding of cues is adequate to account for most experimental results in the experimental literature on category learning. Furthermore, adding third-order cues and beyond does not significantly change the learning or generalization behavior of the network for these experimental tasks.

As an illustration, Figure 4 shows a configural-cue network for classifying geometric patterns varying in size, color, and shape: presentation of a "small white square" causes activation of the input nodes blackened in the figure for single and pair-wise cues. We further assume that these configural-cues obey the same activation and learning rules as do the single features, viz., Equation 3.

Insert Figure 4 about here

A common alternative approach for solving non-linear classification problems is to postulate additional, "hidden units" which connect between the input and output units (Parker, 1986; Rumelhart, Hinton, & Williams, 1986). While these multi-layer networks have considerable power for learning complex discriminations, they require many assumptions regarding their structure (e.g., the basic representation of stimuli and responses, the number and connectivity of hidden units, etc.), their learning rule, and their method for calculating response probabilities. In our limited explorations with these more complex models, we have found none that competes with our configural cue model in fitting experimental data (see, e.g., Gluck, 1991).

Historical Antecedents of the Configural-Cue Model

To include configural cues is hardly a novel move for theories of discrimination learning. Learning theories have traditionally recognized configural learning, since several earlier conditioning experiments demonstrated configural discrimination (Pavlov, 1927) For example, in "negative patterning" (Woodbury, 1943), animals learn to respond to individual stimuli A and B, but to withhold responding to the compound AB. Since both A and B evoke a positive response~~s~~, something other than the summed associative strengths of the component cues must be responsible for the animal learning to inhibit responding to the compound AB. To account for these negative patterning results, yet maintain the essential features of the summation hypothesis, researchers proposed that A and B presented simultaneously generate an additional "configural cue" representing the compound AB (e.g, Hull, 1943; Wagner & Rescorla, 1972, p. 306).

Many analyses have attempted to assess the exact nature of this "unique configural stimulus" and its role in conditioning. Two conceptions of compound AB trials are possible (Kehoe & @@, 1989): presentation of the compound AB may activate either (1) representations of the single cues A and B as well as the configural unit denoted <AB>; or (2) only the representation of the configural unit <AB>. A variety of evidence argues for the former view, that AB presentation causes the arousal of both the components and the configural unit.

~~Further~~ Some characterization of the configural cue is provided by ~~an~~ experimen~~t~~ts by Whitlow and Wagner (1972). Their st~~a~~udies concluded that the configural cue AB is specific to the combination of its individual elements; from the animal's perspective, the configuration is not describable as "two elements," nor as "more intense," nor as "cue A (or B) in the presence of another cue." In a series of studies, Rescorla (1972, 1973) found that configural cues have dynamic associative properties similar to those for single cues. Configural cues can acquire both excitatory and inhibitory associations, their associative strengths summate with those of single cues to determine behavior, configural cues can modify the effectiveness of a given reinforcing event, and their strength can be attenuated by making them irrelevant to the discrimination being trained (see also Kehoe, 19xx; Pearce, 19xx). Thus, our introduction of configural cues into the one-layer network is supported by a considerable history of evidence in animal learning research and theory.[3] Footnote 3

Furthermore, the ability to form configural units and associate them to behavior may rely upon specific brain structures in the hippocampus. Sutherland & Rudy (199@; also Rudy, 19@@; Rudy & Sutherland, 19@@) postulate two parallel learning systems in the mammalian brain: an "Elemental" system that forms associations between single stimulus features and behaviors; and a "Configural" system supported by hippocampal brain structures that brings together several stimulus features into configurations of pairs (or more) and associates these configural units to behaviors. Whereas these two learning systems usually operate together, certain pattern discriminations (e.g., exclusive OR) require reliance on one more than the other system. As ~~direct~~ evidence for their thesis, Sutherland and Rudy (19@@) found that rats that had bilateral lesions in the hippocampus were able to learn simple elemental discriminations but were unable to learn negative-patterning discriminations -- that is, learn to respond to cue A or B alone but to inhibit responding to the compound AB. Sutherland and Rudy (19xx) also use their hypothesis to explain the pattern of results across different learning tasks showing different degrees of sparing versus ~~deficient learning~~ deficits produced by hippocampal lesions in animals. Their hypothesis resembles an earlier one by Wickelgren (19@@) that hippocampal structures are needed to facilitate learning of "higher-order chunks.[4]

Our review of evidence for configural cues was largely confined to studies of discrimination learning with nonhumans. The remainder of this paper tests our network model of configural cue learning in the context of adult humans learning to classify patterns. The stimulus patterns being classified typically involve combinations of three to ten stimulus elements. As noted before, a model utilizing the full power set of all possible subsets of features would rapidly become unwieldy for such expeirments. Consequently, we have arbitrarily limited our configural cue model to use only *pair-wise* conjunctions of the elementary features themselves. While one might argue that the configural cues will have saliences or learning rates that differ from those of the component cues, we will follow parsimony and assume that saliences and learning rates for single-cue and configural-cue units are the same.

In the following sections of the paper, the predictions of this configural cue model will be compared to the data from a variety of representative, critical experiments from the literature on human classification learning. The fit of the configural cue model to the observed data will be compared to that of three alternatives, two of which are network models: (1) our earlier single-cue-only model (Gluck & Bower, 1988a, 1988b), and (2) an extension of the single-cue model proposed by Estes et al. (1989) which uses as inputs ~~only~~ the single cues and the full patterns, so that presentation of stimulus BCD would activate input nodes B, C, D, and the pattern node BCD. We call this the "feature-pattern" model. The fit of the network models to data will be compared to that provided by the exemplar (or "context") model of Medin and Schaffer (1978). It is widely recognized as one of the most successful quantitative theories of classification learning (see, e.g., Nosofsky, 19 ; Smith & Medin, 198@), and serves as a demanding standard to compare against competing models.

COMPARISON TO MODELS OF CATEGORIZATION

The Exemplar Theory

As noted above, traditional theories of categorization presume that people were either using general information about the category abstracted from specific instances (e.g., Franks & Bransford, 196x) or a mixture of category level and specific instance information (e.g., Anderson, Kline, & Beasley, 196x). Medin and Schaffer (1978) proposed a radically different theory -- the "exemplar" or context theory -- in which classification judgments are presumed to arise exclusively on the basis of stored exemplar information. They assumed that specific exemplars experienced during training were stored, and that the subject formed no generalizations or abstractions whatsoever to charactere this array. In their theory, a probe (test) item would be responded to by its ~~first serving~~ retrieving a similar exemplar from memory and then giving the response associated to that exemplar, ~~as a retrieval cue which accesses exemplars similar to the probe~~. The more exemplars of a given class called to mind by a probe, the more likely the subject is to classify the probe into that class.

The context theory predicts that the probability of classifying a test pattern t as a member of category A, P(A | t), is given by

$$P(A \mid t) = \frac{\sum_{a \in A} S(t,a)}{\sum_{a \in A} S(t,a) + \sum_{b \in \hat{A}} S(t,b)} ,$$

where $S(t,a)$ represents the similarity of pattern t to exemplar a, a member of category A, and $\hat{A}$ is the set of all exemplars not in category A. The featural similarity between two exemplars, t and a, is assumed to be the product of the individual-feature similarities, s_i ($0 \le s_i \le 1$), in comparing the individual feature dimensions, i, of the exemplars, viz.,

$$S(t,a) = \prod_{i=1}^{n} s_i .$$

For example, if following conditioning to a red triangle a pigeon were to ~~subject~~ respond~~s~~ 80% as much to a green triangle and 60% as much to a red square, the product rule would expect~~s~~ a proportional response of 48% (80% x 60%) to a green square. The intuition is that the extent to which stimulus feature i evokes a given response depends on the joint context established by the other features which appear together with it -- hence, the name "context model". This similarity rule permits differential weighting of stimulus features (by the s_i's), and implies that the overall similarity of two exemplars will primarily reflect their maximally-dissimilar attribute. With a minimum of processing assumptions, this theory has accounted quantitatively for data from a wide range of classification studies (Medin, 1982; Medin, Altom, Edelseon, & Freko, 1982; Medin, Dewey, & Murphy, 1983; Medin & Schaffer, 1978; Medin & Smith, 1981). It is clearly the "class act" of the set of category models on the contemporary scene (e.g., see Nosofsky, 19xx Estes, 19xx).

The similarity rule. Nosofsky (1984) showed that the similarity rule above is equivalent to the assumption that the similarity between two exemplàrs decays exponentially with their increasing distance in an appropriate psychological space. Considering stimuli composed of separable dimensions, an appropriate distance metric is the Hamming, or city-block metric, in which the distance, $D(x,y)$, between multidimensional patterns x and y is given by

$$D(x,y) = \sum_{i=1}^{n} |x_i - y_i|,$$

where x_i and y_i denote the component values on each of the n stimulus dimensions. A function $f(x)$ will map distance into the multiplicative similarity rule of the context model just in case

$$S(x,y) = f(\sum_{i=1}^{n} |x_i - y_i|) = \prod_{i=1}^{n} s_i .$$

Nosofsky (1984) showed that the above relation can only be satisfied if the function f(•) has an exponential form, viz.,

$$f(z) = e^{-cz}, \quad \text{for} \quad c > 0.$$

Using this identification, the equation for S(x,y) can be written as

$$S(x,y) = e^{(-c\sum_{i=1}^{n}|x_i - y_i|)} = \prod_{i=1}^{n} e^{-c\,|x_i - y_i|},$$

which defines for each dimension i of patterns x and y, a similarity parameter, $s_i = e^{-c\,|x_i - y_i|}$.

Thus, for exemplars composed of separable discrete feature dimensions, the multiplicative similarity rule is equivalent to an exponential decay relationship between pattern distance and psychological similarity. This non-linear relationship implies that the classification of a test stimulus is determined more by ~~this~~ its high similarity to a few exemplars in a category than by its average similarity to all the exemplars. Much of the success of the exemplar model in fitting data can be traced to this implication of its non-linear similarity function (Medin & Smith, 1981).

Surprisingly, the configural-cue model implicitly embodies approximately the same similarity-distance relationship as does Nosofsky's formulation. This equivalence can be seen by noting how the number of overlapping input nodes (similarity) activated by two different patterns changes as a function of their number of overlapping component cues (Hamming distance). Much as the traditional "common elements" theory of generalization (Thorndike, 19xx; Estes, 19xx), the configural-cue network will generalize an association from one to another stimulus pattern in proportion to the number of common input nodes they both activate. The nonlinear relationship between common single features versus common nodes in the configural coding can easily be illustrated. For example, if two quartet patterns share one feature (ABCD,AXYZ), they will have only one active node in common and nine nodes nonoverlapping; if they share two features (ABCD,ABYZ), they will have three nodes in common (two component cues and one configural-cue node) and seven nonoverlapping nodes; if they share three features in common, they will have six active nodes in common (three component cues and three configural-cue nodes). Note how with 1, 2, and 3 common cues, the proportion of overlapping nodes in the configural-cue model increases nonlinearly, as .10 , .30, and .60, respectively. A consequence of this nonlinear similarity metric is that the configural-cue network, like the exemplar model, will judge a 4-element test pattern to be more similar overall to a category of two exemplars with which it shares 1 and 3 single features (for an average of 3.5 nodes in common), than to an alternate category of two

cite equation?

exemplars with which it shares 2 single features each (for an average of 3 nodes in common).

Figure 5 graphically illustrates an example of this similarity-distance relationship for

Insert Figure 5 about here

stimulus patterns composed of five features. As we add pair-wise (doublet) configural nodes, and then triplets in this example, the generalization function more closely approximates an exponential-decay relation. Elsewhere Gluck (1991) has shown that the stimulus representation in the configural-cue network is isomorphic to a special case of Shepard's (1987) theory of stimulus generalization, which derives the exponential-decay relationship between similarity and distance as an optimal strategy in the face of uncertain "consequential regions" around a trained stimulus.

The fact that the configural-cue model implies a ~~y~~near-exponential generalization gradient partly explains why we have a changed ~~in~~ our "response rule~~."~~ from (Gluck + Bower, 198X). With ~~In~~ earlier work ~~with~~ the single-cue version of the model, we had converted category activations into response tendencies by an exponential transform, i.e., $v_i = e^{\Theta A_i}$. The configural cue model puts the near-exponential response tendencies of overlapping patterns into the stimulus coding (allowing a simple ratio response rule) rather than into an arbitrary transformation of activation levels into response selection (a similar point was made by Nosafsky, 199x).

FITTING THE CONFIGURAL CUE MODEL

Here goes new analysis of Gluck & Bower '88 data

The Medin and Schaffer (1978) Results

Medin & Schaffer (1978) presented data from several experiments which consistently favored the context model in comparison to independent-cue models, ~~which~~ include~~s~~ing the prototype-averaging model. Independent-cue models compute the predictive power, or association, of each cue dimension separately. The single-cue network model of Gluck and Bower (1988a) is ~~similar to~~ an independent-cue model~~s~~ since it independently combines information about individual cue-outcome associations to determine the network's response to a stimulus pattern. The configural cue model differs ~~from independent-cue models~~ in that cue-outcome associations are learned between cue combinations and responses, and so cues are not *independent* of each other. Rather, activation patterns across output responses depend strongly on a sensitivity to all cues co-present during training. In the following, we will compare the predictions of the configural-cue model to the data offered in support of the exemplar model. We will also consider the predictions of the "feature-pattern" network model of Estes et al. (1989) which extends the single-cue representation by adding additional nodes representing the presence or absence of entire patterns. While such full-pattern nodes enable such a model to learn non-linearly separable classifications, this model nonetheless embodies the same similarity function as the single-cue model, viz., similarity simply equals the proportion of shared component features.

We turn now to comparing these three models' abilities to account for the data from Medin & Schaffer (1978).

Experiment 1 ~~*of*~~*:* In Experiment 1 by Medin & Schaffer, subjects learned to classify six patterns into two equal categories. The training patterns were chosen so that the context model and independent-cue models would make different predictions concerning each pattern's relative learning difficulty. The ~~Each~~ patterns consisted of geometric shapes varying in binary valued dimensions such as shape, size, and color; ~~and we denote~~ we adopt the standard convention of denoting the two values of each attribute dimension as 1 and 0. As Table 2 illustrates, ~~a~~ the different cue values on ~~any~~ each dimension

----

Insert Table 2 about here

----

were ~~was~~ ~~so~~ labelled arbitrarily so that a value of 1 was more diagnostic of Category A, ~~at~~ and the ~~a~~ the cue-value labelled *0* was more diagnostic of Category B. The diagnosticity of the 1-0 values held within every dimension. Therefore, models which add up the diagnostic evidence from ~~each~~ the cues independently will assign equal weights to the four dimensions. One implication of this view is that the category prototypes *1111* and *0000* in Table 2, should be equally difficult to learn. ¶ The context model, however, considers the similarity of exemplars within a category versus between categories as the major determinant of item difficulty. In these terms, the Category B prototype, *0000*, has one highly similar pattern in its own category (*0100*) ~~which differs from it on only one dimension~~ but another pattern (1011) that is quite dissimilar. In contrast, the Category A prototype, *1111*, has two moderately similar patterns (2 of 4 features in common), but also has a highly similar pattern (*1011*) in the opposing training category. This balance of within-class versus between-class similarities causes the context model (with its nonlinear similarity rule) to predict that pattern *1111* will be harder to learn than pattern *0000*.

After training, subjects were presented with a series of transfer test patterns (bottom rows of Table 2) designed to discriminate between the models. Since during training the cues within each dimension were equally diagnostic of category membership, independent-cue models will assign equal associative strengths to the cues of the four dimensions. ~~Therefore,~~ According to such models, then, the degree of association of an instance to Category A ~~is~~ should be given simply by its total number of 1-values and its associative strength towards Category B is given by its number of 0-values. Each transfer pair (row of Table 2) compares a Category A item containing three 1's with a Category B item consisting of three 0's. Thus, independent-cue models would classify each item of a test pair into its respective category with equal strength, because these models are only sensitive to the number of 1's versus 0's.

The context model, however, makes different predictions. The transfer items *b* 1, *b* 2, *b* 3 are highly similar to one A and one B pattern, whereas the *a* 1, *a* 2, *a* 3 transfer patterns are highly similar to two A exemplars and no B exemplars. Thus, the context model predicts that *a* 1, *a* 2, *a* 3 should be classified more accurately as A's than their respective *b* 1, *b* 2, *b* 3 mates are classified as Bs.

Turning to the results, Medin & Schaffer found that the learning and transfer data from Experiment 1 supported the context model's predictions. During training, subjects made more errors on the *1111* prototype than on the *0000* prototype, and they gave more confident A responses to the column A transfer patterns than they gave B responses to the column B patterns.

The three network models (single-cue, configural-cue, and feature-pattern models) were tested in a simulation of the training ~~of~~ used in Medin + Schaffer's Experiment #1. In each case, the networks were trained through 20 random-order presentations (epochs) of the six training exemplars since 20 was the maximum number of epochs that actual subjects received. For each model, the learning rate parameter, β, was selected to minimize the mean squared error between the predicted and observed cumulative errors reported for each training pattern. Each network consisted of two output nodes, one representing Category A, the other Category B. An output (category) node was reinforced with λ=1 if a particular pattern was a member of that category, and with λ=0 if the pattern was ~~not~~ a member of ~~that~~ the alternative category. The single-cue model had eight input nodes, two for each of the feature values on the four dimensions. The feature-pattern model included six additional input nodes corresponding to the presence of each complete stimulus pattern. The configural-cue model expanded the single-cue representation by the addition of 24 nodes corresponding to the pair-wise combinations of cues that occurred in the six training patterns.

The expected probability that a subject would make a particular classificatory response was calculated by dividing the activation of that category's output node by the summed activation of all output nodes. To ensure that this mapping resulted in an expected probability between 0 and 1, negative output activations (which were quite small, infrequent and generally only occurred very early in training) were taken to be 0.[4] The models' predictions for the total cumulative errors for each training pattern was computed as the sum, across all presentation trials, of the probability of a classification error for that pattern. The data that Medin + Schaffer reported for Experiment 1 are neither sufficiently detailed nor reliable to evaluate the models quantitatively. In place of that, we will evaluate the models by their ability to predict qualitative, ordinal features of the data.

To give our conclusion first, these qualitative comparisons showed that the configural-cue model predicted the data more accurately than either the single-cue or the feature-pattern models. The predictive advantage was evident for both the learning data and the transfer test data. Of course, the single-cue model cannot fit the learning data well because it cannot learn to criterion the nonlinearly separable classification of Table 2. However, the three models correctly rank-order the difficulty (errors) of the six patterns during training, with patterns $b2$ and $a2$ being hardest to learn, and $a1$ and $b1$ being easiest. The rank-order correlation between the observed and predicted average numbers of errors for the three models was 1.0, 0.94, and 0.94 for the configural-cue, single-cue, and feature-pattern models, respectively. The critical comparison concerned the difficulty of learning patterns $a1$ and $b1$. Here, the configural-cue model correctly predicted that $a1$ would be harder to learn than $b1$ (2.62 versus 1.12 mean predicted cumulative errors), whereas the alternative models predicted that $a1$ and $b1$ would be about equally easy to learn (2.96 versus 3.09 predicted cumulative errors for the single-cue model, and 1.29 versus 1.33 errors for the feature-pattern model on patterns $a1$ and $b1$, respectively).

Turning to the pairs of test patterns (see Table 2), Medin & Schaffer reported the average scaled confidence with which subjects assigned each pattern to category A or B. Taking the difference in confidence with which test patterns ai and bi are assigned to categories A and B, respectively, the differences were positive for all pairs. That is, subjects considered each ai pattern to be a better example of the A category than its bi pattern mate was of the B category. Importantly, these orderings are exactly as predicted by the configural-cue network and the context model, whereas the single-cue and feature-pattern networks expect no differences whatsoever in these paired comparisons.

In summary, the learning and transfer data of Experiment 1 support the configural-cue model in preference to independent-cue or prototype models, including the single-cue model alone or ~~that~~ when single cues are augmented by ~~a~~ unique nodes for storing complete patterns.

Experiments 2 and 3 of Medin & Schaffer: As noted, Experiment 1 taught subjects categories that were not linearly separable. Experiments 2 and 3 in Medin & Schaffer were designed to quantitatively assess the performances of an independent-cue model and the context model for a linearly separable classification task. Both experiments had the same structure: Experiment 2 used geometric forms as stimuli, whereas Experiment 3 employed schematic Brunswik faces varying in type of eyes, hair, nose, forehead, and so on. Table 3 presents the schematic experimental design. As before, cue values are labelled so that value 1 on any dimension appeared more often with Category A whereas value 0 appeared more often with Category B. The prototypes for categories A and B were *1111* and *0000*, respectively. The categories were linearly separable in that a majority of 1's on dimensions 1, 3 and 4 indicates Category A; otherwise, the pattern is classified in Category B.

Insert Table 3 about here

The main difference between the predictions of the independent-cue model and the context model concerns the relative difficulty of learning patterns *a*1 and *a*2. Pattern *a*1 contains more characteristic values (1's) indicating Category A than does stimulus *a*2 and is also closer to the category prototype *1111*; ~~so~~ therefore, independent-cue models predict that pattern *a*1 should be easier to learn than pattern *a*2. The exemplar-context model, however, makes the opposite prediction. Pattern $a\,2$ is highly similar to two Category A stimuli ($a\,1$ and $a\,3$) and not highly similar to any Category B stimuli. In contrast, Pattern $a\,1$~~, however,~~ is highly similar to only one Category A stimulus ($a\,2$), but is also highly similar to two Category B stimuli ($b\,1$ and $b\,2$). Thus, according to the context model, the combination of more within-category similarity with less between-category similarity should make pattern $a\,2$ easier to learn than pattern $a\,1$. Confirming this prediction, subjects in Experiments 2 and 3 committed fewer errors in learning stimulus $a\,2$ than stimulus $a\,1$ (Table 3). The context model was also supported by its ability to predict the proportions of times subjects categorized the test probes as As and Bs.

We now examine the networks' ability to fit the data from Experiment 2. All three network models were trained on the training patterns of Experiment 2 for 16 epochs (the maximum number of epochs seen by subjects in Experiments #2). Because the structures of both experiments were identical, the qualitative predictions concerning stimulus similarity are the same for both experiments.

First, we consider the difficulty subjects experienced in learning the nine different training patterns. The learning rate parameter, β , was selected for each model to best fit the transfer test data to be reported below; this same β was then used to simulate performance during the training phase. Each model predicts the expected cumulative errors during training on each of the nine patterns in Table 3. The rank-order correlation between the predicted and observed average errors was ? .98 for the configural-cue model, .82 for the feature-pattern model, and .80 for the single-cue network model. Thus, the configural-cue model predicts the relative difficulty of learning the nine patterns nearly perfectly, and better than the two alternative network models. The context model was not fit to the course of learning by

with 9 elements ?

Medin & Schaffer, since at that time no realistic gradual learning assumptions (e.g., later offered such learning assumptions.) regarding faulty storage) had yet been offered (~~but see~~ Estes, 198@)

We turn now to the transfer test data. Medin & Schaffer reported the proportion of category-A choices for the 16 patterns presented during the transfer tests: seven novel patterns (*n* 1 to *n* 7 in Table 3) along with the nine training patterns. The fits of the four models to these data are shown in Table 3. In general, each of the models is capturing the overall ordering of choice proportions among the 16 patterns, except for pattern *a* 3 which is overpredicted by all the models for some unknown reason. However, the context model's predictions required the estimation of four parameters (the similarities of the two values on the four stimulus dimensions). Its predictions correlate .97 (rank-order) with the observed proportions. On the other hand, the similar correlations for the configural-cue, feature-pattern, and single-cue network models are .97, .88, and .93, respectively. ~~These latter three correlations are noteworthy in two respects~~. ~~First,~~ The ~~configural-cue model fits the data somewhat better than the other two. Thus, even with linearly-separable classifications, the single-cue network model does not predict the data as well as the configural cue model.~~ ~~Second, the configural-cue~~ three networks model s required ~~the estimation of only a single parameter (the learning rate, β) whereas the context model required the estimation of four parameters based on only the 16 proportions in Table 3. For this reason, there is considerably more leeway for the~~ network ~~configural-cue~~ ~~model~~ s ~~to misfit these data than there is for the context model. Consequently, the close fit of the configural-cue model implies that the data points are indeed being constrained by one another in much the way the model envisions.~~

said just below

Discussion of the Medin & Schaffer Tests

Both the context and the configural-cue models provided good quantitative fits to results of Medin and Schaffer, whereas the feature-pattern and single-cue models fared less well. However, the network model offers several advantages over the context theory. First, the configural-cue network fit both the learning and the transfer data after estimating just one free parameter, the learning rate β. In contrast, the original context model only fit the transfer data and for that required the estimation of four similarity parameters. Second, while the multiplicative similarity rule (or its exponential version) is a bald assumption in the context theory, it is an emergent property inherent in the stimulus representation of the configural-cue network. Third, the network model provides a trial-by-trial account of learning, while the original context model only provides a best-fitting set of similarity parameters valid at some specific stage of training, e.g., at completion of training. On these bases, then, the configural-cue model would be favored over the context model. However, one set of results hardly settles the question of the better-fitting model. We require comparisons of the models' fits to other critical results, to which we now proceed.

@ GB: Do you think the F-P Effect goes best here?

THE FEATURE-POSITIVE EFFECT

One elementary observation that seems to favor the configural cue model over the exemplar model is the Feature-Positive effect (Jenkins and Sainsbury, 1970; Sainsbury, 1971;1973; Newman, Wolf, and Hearst, 1980). The effect arises in a simple discrimination learning task, comparing the learning of the discrimination AB+ versus A-, to the learning of the complementary discrimination, A+ versus AB-. The former task has the positive (reinforced) stimulus marked by the presence of a positive sign, cue B; the latter task has the negative (nonreinforced) stimulus marked by presence of this added sign. The invariable result, found both with rats, pigeons, and human adults, is that the Feature-Positive task is far easier to learn than the Feature-Negative task.

This difficulty-ordering is difficult for an exemplar model to explain. In either case the AB and A patterns are repeatedly stored with their respective correct responses and the nature of the similarity calculations should not depend on which stimulus is positive and which is negative. Even should the exemplar model be augmented with selective attention to the extra feature, there is still no reason for that theory to expect a difference depending on whether that feature is positive or negative.

The Feature-Positive advantage ~~falls out of~~ is implied by the inherent characterization of the two discrimination tasks by the configural cue network model. The three sensory units are A, B, and (AB), and we assume that these begin with no associations to the reference response. In the Feature-Positive task, during the course of AB+ versus A- learning, the association of the B and AB units increase to .50 whereas the A unit's association rises some initially then returns to zero; the eventual result is an activation of 1.0 to the AB+ compound and 0 to the single cue A-. In contrast, in the Feature-Negative task, the association of cue A must increase to a strength of 1.0, while the associations of cues B and (AB) are to decrease to strengths of -.50. Comparing the two cases in the model, the Feature-Positive discrimination is learned more readily simply because the amount of change of associative strengths of the three sensory units (namely 1.0) is less in that case than in the Feature-Negative case (where changes must add to 2.0). In one simulation of the configural cue model, for example, after ten training epochs, the mean squared error was .035 for the Feature-Positive task and nearly four times as large (.130) for the Feature-Negative task. Moreover, in the Feature-Positive task, the model shows an initial increase followed by a decrease in responding to A-, an acquisition pattern invariably observed in such discrimination learning.

We conclude that the configural cue model provides a viable explanation of the Feature-Positive advantage, an effect which would seem difficult for the exemplar model to explain.

LINEAR SEPARABILITY AND CLASSIFICATION DIFFICULTY

Historically, the inability of one-layer networks to learn non-linearly-separable classifications has been a major limitation to their general application. But we may ask whether linear separability an important variable controlling human classification learning. Experiments by Medin and Schwanenflugel (1981) provide relevant comparisons. They contrasted performance of subjects learning pairs of classification tasks, one which was linearly separable and another, closely matched for similarity, which was not. The context model of Medin & Schaffer predicts that exemplar similarity, not linear separability of the two classes, will be the major determinant of classification difficulty. Highly similar exemplars within a category should facilitate classification learning, whereas similar exemplars belonging to different categories should retard learning. For two of their experiments (#3 and #4), Medin and Schwaneflugel constructed the exemplars so that exemplar-similarity would favor the non-linearly separable over the linearly-separable classification. In these cases, then, the exemplar-context model predicts that the non-linearly separable classification would be easier to learn. We will begin by examining the three dimensional task used in their Experiment

#4, since its stimulus design is easier to explain. We will then turn to their Experiment #3 in which subjects learned a four-dimensional problem, for which more detailed data were reported for us to simulate.

The three-dimensional classifications can be visually represented (see Figure 6) using

Insert Figure 6 about here

the eight corners of a cube. Figures 6A and 6B are the three-dimensional linearly-separable (LS) and nonlinearly-separable (NLS), respectively, classifications studied in Experiment #4. This graphical representation makes apparent the linear separability of the LS classification; one can visualize a plane slicing the cube, separating Category A exemplars (black corners) from Category B exemplars (white corners) in Figure 6A. No such plane exists for the NLS classification in Figure 6B.

As mentioned, the similarity rule of the context model predicts that this particular LS task will be more difficult to learn than the NLS task. To understand this prediction, 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 to each white dot in Figure 6, and take the average distance. The Hamming distance between two exemplars is the minimal number of edges traversed in moving from one stimulus corner of the cube to another. For the LS task in Figure 6A, six-ninths of the between-category distances are one edge apart and three-ninths of them ~~the distance~~s are three edges apart, yielding an average between-category distance of 5/3. For the NLS task in Figure 6B, four-ninths of the between-category instances are one edge apart, four-ninths are two edges apart, and one-ninth are three edges apart, for an average between-category distance of 5/3. Calculation of the average within-category distances proceeds similarly, yielding the value of 2 for both classifications. Thus, the LS and NLS stimuli ~~have been~~ were cleverly constructed so that both tasks have identical average within-category distances (of 2), and identical between-category distances (5/3).

What differs between the two tasks is the distribution of between- and within-category distances. The histogram in Figure 6A shows that the linearly separable task is composed of relatively more "close" (Distance=1) and fewer "far" (Distance=3) relations, whereas the non-linearly-separable task (Figure6B) has a more even distribution of "close", "medium", and "far" between-category distances. These unequal intra-item distance distributions have important implications when there is a non-linear mapping from Hamming distance to similarity. Both the configural-cue model and the context model exaggerate the psychological similarity of close exemplars. The greater proportion of these close between-category distances in the LS (compared to the NLS) classification should increase the confusion between the two categories; thus, both models predict that the LS classification in this comparison (Table 6) should be more difficult to learn. A similar analysis of within-category distances indicates the presence of fewer close distances in the LS task compared to the NLS task; this also suggests that the LS task should be more difficult.[5]

Although we have used the three-dimensional task to illustrate the nature of the LS and NLS tasks and the rationale underlying their design, Medin and Schwanenflugel actually reported in most detail the results of a four-dimensional task (their Experiment #3) which we will compare to the predictions of the alternative models. Table 4 schematizes the two

Insert Table 4 about here

groups of six stimulus patterns that their college students learned to classify into two categories. The stimuli were photographs of human faces which varied along four binary dimensions: hair color, shirt color, smile type, and hair length. Eighteen different photographs were used to exemplify each of the six stimulus types, so over 108 trials subjects never saw the same face twice. The LS classification at the top of Table 4 is linearly separable whereas the NLS classification at the bottom is not. To recognize the linear separability of the LS classification, note that the number of 1's in dimensions 1,3, and 4 equal two for any category A stimulus, but is less than two for any category B stimulus. No such linear combination of feature values will perfectly separate the two classes of patterns in the lower, NLS classification.

Mark's new data

While a geometric representation of these four-dimensional classifications is not feasible, the distribution of intra-exemplar distances can be analyzed as before (see Figure 7). Again the LS and NLS classifications were cleverly designed to have identical between-category average Hamming distances (20/9) and identical within-category average distances (8/3); but the distribution of these distances varies between the two tasks. Compared to the NLS task, the LS task had more close (distance=1) between-category relations and more close (distance=2) within-category relations.

Insert Figure 7 about here

Figure 8A shows the main result from Experiment #3, namely, the error (learning) curve throughout training. As is obvious, subjects found this LS classification to be harder to learn than its matched NLS classification. This reversal of what most other models expect is strong testament to the predictive power of exemplar models which have an exponential generalization function. Note that an exemplar model with a proportional (linear) similarity rule would not predict ~~the~~ these results.

Insert Figure 8 about here

Comparing Alternative Network Models

To investigate whether the configural-cue model could also fit these results, we trained that model on the stimulus structure shown in Figure 2 (see also Gluck, 1991). Figure 4B shows the configural-cue model's predictions for this experiment. As did the context-model, the configural-cue model correctly predicts that subjects will find this LS task more difficult than this NLS task.

We also simulated the experiment with three alternative network models: the single-cue mode, the "feature-pattern" model, and several multi-layer "back-propagation" network models (Rumelhart, Hinton, & Williams, 1986) which are capable of learning nonlinearly-separable classifications.

Applying these models to the stimulus design of Table 4, we find that, as expected, the single-cue network learns the linearly separable task but never fully learns the non-linearly separable task; thus, it expects the NLS task to be more difficult. Similarly, the feature-pattern model also incorrectly predicts that subjects should find the linearly separable task easier to learn.

To evaluate multi-layer networks, we adopted the same stimulus representation on the input and output nodes as the single-cue model. Our multilayer networks had one additional layer of "hidden" units, fully connected to nodes in both the input and output layers. To train the network we applied the "backpropagation" rule, a generalized version of the LMS learning algorithm (Parker, 1986; Rumelhart, Hinton, & Williams, 1986; Werbos, 1984). Regardless of the number of hidden units employed (2, 8, or 16), this multi-layer backpropagation network incorrectly predicted that subjects should find the linearly separable task easier to learn. Figure 8C shows illustrative learning curves produced by a two-layer network with two hidden units.

The above theoretical comparisons help us judge the diagnostic value of the Medin and Schwanenflugel results. Their counterintuitive result, that a nonlinearly-separable classification can be arranged to be learned more easily than a linearly-separable one, challenges the single-cue, prototype, and feature-pattern models since no parameter values will suffice to make those models fit ~~the~~ such an outcome. The situation for multi-layer network models is less certain; while our straight-forward architectures failed to predict the data, the class of multilayer models is so huge and unconstrained that we cannot claim to have searched the entire class for a successful contender. Our search to date has been unsuccessful. In contrast to these failures, the configural-cue network and the context model both predict the qualitative ordering of the LS and NLS curves in Medin and Schwanenflugel's results regardless of parameter values. Such parameter-free predictions provide strong support for the configural-cue and context models against other contenders.

@@ WE HAVE A REPLICATION OF THE LS/NLS TASK WITH EVEN STRONGER EFFECT (maybe generalization data, too). POSSIBLE ALSO TO FIT LEARNING CURVE. SHOULD WE INCLUDE? OR SAVE FOR LATER? -mg

EXEMPLAR SIMILARITY AND RULE ABSTRACTION

By tradition, ~~We have noted~~ three broad classes of models for learning classifications may be distinguished. First are the prototype abstraction models (Franks & Bransford, 1974; Posner & Keale, 1968) which infer an average or modal description of exemplars of a class by recording features independently. Second are exemplar-storage models (Medin & Schaffer, 1978) which perform no abstraction but which classify a transfer stimulus according to its relative similarity to full exemplars of each class stored in memory. Third are power-set frequency models (Hayes-Roth & Hayes-Roth, 1977; Reitman & Bower, 1973) which accumulate category diagnostic frequencies for all possible combinations of features (singlets, doublets, triplets, etc.) as these appear in the training patterns, and which classify a test pattern according to the diagnostic weights of the several feature-combinations it contains. Each of these general models have several variants. Our configural-cue model is a variant of a power-set frequency model in which we have restricted the diagnostic strengths to be accumulated only for single features and pairs of features and have used the LMS learning rule. The ACT model of Anderson, Kline, and Beasley (1979) is another example of a power-set model, but one in which any and all conjunctive generalizations describing a subset of exemplars may be learned and then strengthened depending on their diagnostic success.

While the accumulated evidence allows the rejection of the prototype abstraction model in critical experiments, the exemplar-storage and power-set models have been closely parallel in their account of most results. An important set of experiments by Elio and Anderson (1981) attempted to differentiate between the latter two classes of models, specifically between the Medin and Schaffer exemplar model and the ACT model. Elio and Anderson's experiments manipulated the likelihood that subjects could form category abstractions or rules; they achieved this end while holding constant (between different groups) the similarity of transfer items to the set of training patterns.

The method Elio and Anderson used to arrange these conditions was simple and elegant. An experimental group of subjects learned to classify a set of five-featured stimuli which included in the same category ~~the~~ two patterns denoted generically as ABFGH and ABIJK; in this case, an AB--- rule is possible. They were then tested on a pattern such as ABCDE. Note that this test pattern overlaps in two features with the two training exemplars and with the more general, AB--- rule. For comparison, a control group learned a category that ~~stimuli~~ includ~~ing~~ed the patterns ABFGH and IJKDE, before receiving the ABCDE test pattern. For these two patterns, no rule abstraction is possible; still, each of these training patterns shares in exactly two items with the test pattern ABCDE, the same degree of overlap as for the experimental subjects. In terms of the number of shared elements, then, the experimental and control conditions are equated. Consequently, a simple exemplar-storage model would expect no difference in response of the two groups to the transfer stimulus. On the other hand, the ACT model expects the experimental subjects to learn the more general AB--- rule and use it to classify the transfer pattern more accurately than will the control subjects.

The data provided slightly more support for the ACT model than for the exemplar model. Although the data were qualitatively in accord with the ACT model, certain quantitative aspects of the data were not well explained (see Anderson, 1990, p. 113). Because of the critical role these experiments have played in the theoretical controversies regarding classfication models, we wished to compare the fit of the configural cue model against that of the ACT model to these results. We now turn to describing the details of the experiments to be simulated.

In their Experiment #1, Elio and Anderson's subjects studied five-feature descriptions of 16 fictitious people who belonged to one of two social clubs. ~~The training set consisted of 16 person-descriptions, half of which were identified as belonging to one social club, and half to another club.~~ The five attributes along which the people varied were their religious affiliation, hobby, job, education level, and marital status. Each attribute could take on one of four different values. A sample training item ~~was~~ might be, "One member of the Dolphin Club is a Baptist, plays golf, works for the government, is college educated, and is single." ~~As before~~, we will use five-digit vectors, composed of the numbers 1 through 4, to represent the four possible feature values, one for each of the five attribute dimensions. In the *Rule* condition, @give #@ pairs of exemplars in a category had the same value (overlapped) on two of the five, allowing subjects to form two-element generalizations; other exemplar pairs overlapped on three of the attributes, permitting three-element rules. The exemplars learned by subjects in the *Control* condition did not contain such generalizations.

Training was carried out to a criterion of one perfect cycle through the 16 training patterns. Following training, subjects in both conditions classified two sets of 16 transfer patterns, a three-overlap and a two-overlap set. In both the rule and control conditions, each pattern in the *two-overlap* or *three-overlap* transfer set overlapped with two of the original training exemplars on two or three of the five dimensions. In the rule condition the two- and three-overlap transfer stimuli overlapped with training exemplars on the shared critical two or three dimensions of the potential rule. For example, subjects in the rule condition might study *11213* and *11312*, and then be tested with the three-overlap transfer item *11111*, whose values overlap with those for the first, second, and fourth dimensions of both patterns. In the control condition the three-overlap transfer stimuli also overlapped on two or three dimensions with two training patterns. However, this overlap was not consistent with any potential rule.

Classification accuracy of the transfer patterns was compared for the four possible conditions in a 2x2 design: rule training set and three-overlap transfer test (denoted "r3"); rule training set and two-overlap transfer test (denoted "r2"); control training and three-overlap transfer (denoted "c3"); control training and two-overlap transfer (denoted "c2"). The experiment thus compares the influence on transfer responding of general rules versus exemplar similarity. For example, if people consider exemplar similarity alone when they make category judgements, their classification response should be more accurate on the three-overlap than the two-overlap transfer items but not differ between the control and rule conditions (*i.e.* r3,c3 > r2,c2). If subjects make use of rules alone, however, then accuracy should be better in the rule condition than the control condition, but should not differ between three-overlap and two-overlap transfer items (*i.e.* r3,r2 > c3,c2).

The mean accuracy ratings observed by Elio and Anderson for the four conditions (see Table 5) shows that within each rule or control condition, people classified the three-

---

Insert Table 5 about here

---

overlap transfer items more accurately than the two-overlap items. This presumably reflects the influence of exemplar similarity on transfer performance. Furthermore, with overlap held constant, subjects performed more accurately on transfer items when they coincided with potential rules; we express this relationship as r3 > c3 and r2 > c2. The magnitude of these two main effects -- rules ~~vs.~~ and similarity -- are roughly equal in these average data.

Turning to the simulation models, we note that the rules defining club membership make the clubs linearly separable. We might therefore expect that even the single-cue network model ~~w~~should predict these two main findings. For example, a pair of exemplars that overlap in the rule training set should cause the weights on the shared features to acquire a relatively large associative strength to the category at the expense of the other dimensions. In the control condition, however, exemplars do not share features often, so only small weights would accrue to these features. Thus, we might expect the network's response to a test pattern to be stronger in the rule condition, because the network's output activation would reflect the sum of several larger weights. Similarly, classification should be stronger with the three-overlap than the two-overlap transfer items, because the output sum for the three-overlap stimuli would simply include a greater proportion of the larger weights.

To check these intuitions we trained the single-cue network using the procedure from Elio and Anderson's Experiment #1. The network consisted of one input node for the four possible cue values on each of the five stimulus dimensions (20 total), and two output nodes representing each of the two categories. The number of training epochs was set equal to the average number of epochs required by Elio and Anderson's subjects to reach learning criterion (viz., 13 for rule and 16 for control subjects). We estimated the learning rate (at β*eq* .008) so as to minimize the sum squared error between the network's predicted choice accuracies and those observed for the 16 transfer patterns (Elio & Anderson, Table A1, p. 416).

As shown in Table 5, the single-cue network model correctly forecasts that the three-overlap transfer items would be easier to classify than the two-overlap patterns for both the rule and the control conditions (i.e., r3 > r2 and c3 > c2). However, this model erroneously predicts that subjects should about equally accurate in the control condition compared to the rule condition (c3 ≈ r3; c2 ≈ r2). Thus, contrary to intuitions, the single-cue network does not correctly predict that the existence of potential rules increases subjects' accuracy in this task. Later we will examine what factors underlie these failed intuitions.

The configural-cue network, however, provides more accurate predictions. Adding input nodes representing pair-wise cue combinations and estimating a new β (at .00@), the configural-cue model correctly predicts the two main results (Table 5). That is, the configural-cue model correctly expects that potential rules as well as exemplar similarities would both improve classification accuracy.

We will briefly note that the feature-pattern model (of Estes et al., 1989) fares poorly with these data because its predictions were identical to those of the single-cue network and suffered from the same shortcomings. The erroneous predictions of the component-cue model are unaffected by the addition of whole-pattern nodes for two reasons: first, the weights corresponding to each single pattern node are overwhelmed by the weights associated with the 20 single-cue input nodes; second, the addition of pattern information provides no information whatsoever about the existence of either potential rules or pattern similarities.

To understand the failure of the single-cue network model to capitalize on the potential rules, a closer examination of the category structure is required. Although the transfer patterns overlapped intentionally with specific training pairs designed to form three-overlap and two-overlap rules, a number of unintended overlaps of the test stimuli with other training exemplars also arose. These unintended overlaps appear to have caused the mispredictions of the single-cue network. The configural-cue model is less sensitive to these spurious overlaps because the addition of doublet nodes increases the network's relative sensitivity to multiple overlaps.

Elio and Anderson noted these unintentional overlaps between training and transfer test patterns. To assess their impact, they computed a mean similarity score for each of their transfer items using a multiplicative rule similar to that employed by Medin and Schaffer (1978). The average of the similarities for the patterns in each condition are shown in Table 5 (last column). The ordering of these similarity scores parallels the two main results of Experiment #1. That is, the similarity ratings for both the three-overlap and two-overlap rule conditions are higher than for the three-overlap and two-overlap control conditions (i.e., r3 similarity > c3 similarity, and r2 similarity > c2 similarity). Thus, the Medin and Schaffer (1978) context model may be expected to predict these data nearly as well as the configural-cue network model, viz., r3 > c3 and r2 > c2 (see also Nosofsky, 19@@).

delete the above

Experiment #3 of Elio and Anderson: Elio and Anderson's Experiment #3 was designed to overcome some limitations in the design of their first experiment. The new study had essentially the same design as before except that to reduce unintended overlaps each training set consisted of exemplars with five possible cue values on each of the five dimensions. Both training sets had separate transfer tests so that the mean similarity between training and transfer patterns could be made as equivalent as possible across both the rule and control conditions. Only two conditions were studied, a rule training set paired with rule transfer test items, and a control training set with associated control transfer test patterns (stimulus overlap was not varied). The results showed that subjects' transfer performance was more accurate following study of exemplars for which rules were available compared to the control study set.

We trained the single-cue, feature-pattern, and configural-cue network models in a simulation of their Experiment #3. The input nodes were those corresponding to the possible combinations of five cue values on five stimulus dimensions relevant to each model. The control and rule conditions were both trained for 11 epochs, approximately the number required by subjects to reach criterion. ~~As before~~, the learning rate, β, was chosen to minimize the sum of squared errors between the predicted and observed transfer test accuracies (Elio & Anderson, Table A3, p. 417). With this procedure, we found that all three network models ~~now~~ correctly predicted the main advantage of having rules available. For the configural-cue model, the predicted mean accuracy in the rule condition is 0.79 (compared to an observed value of .79); the predicted mean accuracy in the control condition is 0.70 (compared to an observed value of .72). Both the single and feature-pattern models predict mean accuracies of .78 in the rule condition and .69 in the control condition.

The fit of the network models to the Elio and Anderson results is encouraging because they are often cited as demonstrating the role played by rule abstraction in category learning. But note that the rules in these cases are expressible simply as the sharing of feature-conjunctions across several training exemplars. The sharing of feature conjunctions implies that the associative strength of the connection between that conjunction-node and the category will often be ~~incremented~~ strengthened, thus enabling it to overshadow and compete successfully with single-cues or other feature combinations.

The ACT model of Anderson et al. (1979) differs from the configural-cue model in several aspects. As applied to an experiment similar to those of Elio and Anderson, for example, the ACT model could learn ~~three-element rules~~ rules based on three elements, of the form (ABC--)→Club 1 as well as the more general rules (AB---)→1, (A-C--)→1, and (-BC--)→1. In contrast, when restricted to only pairwise conjuncts, the configural-cue model represents a three-element generalization simply in terms of large weights on the three conjunctive nodes for AB, AC, and BC. In fitting the Elio and Anderson results, this restriction seems to have been immaterial.

~~The~~ One major difference from the configural cue model is that the ACT model ~~allows for~~ permits any Boolean or predicate-calculus function of the stimulus features to serve as the left-hand side of a rule. For example, a possible ACT rule (see Anderson et al., 1979) might be "If a person is a Catholic and a tennis-player but is not a day laborer or divorced, then he is probably a Club 1 member." [A Boolean expression is one formed by conjunctions, disjunctions, and negations of any length of elementary features.] However, when ~~one has~~ the stimulus domain comprises a small set of dimensional values, then negations are equivalent to a disjunction of alternative features (e.g., *not* brown eyes implies they are either blue or hazel). Moreover, a rule involving the disjunction of features is equivalent to a simple listing of element-to-category associations. The result of these considerations is the view that any rule expressible as a Boolean function of elementary stimulus features can be captured and expressed as well by a disjunctive listing of associations of configural (conjunctive) cues in our model. In fact, a theorem in Boolean algebra (see Suppes, 19@@) attests that any Boolean function can be expressed as the disjunction of a collection of conjuncts.

nearly? conjuncts only? ?

The ACT model has potentially more expressive power than the configural-cue network simply because it can accomodate any predicate calculus (or propositional) expression on the lefthand side of a rule. Thus, for example, in a rule for classifying a printed letter as a capital letter E, the lefthand side might be "If the left ends of three parallel, horizontal lines touch a vertical line with no figures closing them on the right edge, then the figure is an E." Note that several predicates here are relational (left end, parallel, touch) and reflect a rich perceptual base. It is this relational perceptual base that gives such rules far more descriptive power than models (such as the configural-cue network) based only on Boolean functions of presence/absence of featural elements. In defense of the configural-cue model, however, we would point out that the overwhelming majority of studies of the learning of artificial categories have been carried out within the latter, restricted domain. So we will continue to test the configural-cue model as appropriate for that well-researched domain of classification problems.

@@@@@@@@@@@

@NOTES:

SHOULD WE DELETE EXPERIMENT #3 SINCE IT DOESN'T DISTINGUISH AMONG THE

NETWORK MODELS????

THE ANDERSON CHAPTER ALSO ONLY ANALYZES EXPERIMENT #1

(BUT DOES NOT TOUCH

ON THE SPURIOUS OVERLAP/SIMILARITY PROBLEM WHICH I THINK

IS IMPORTANT)

NOSOFSKY DOES NOT ANALYZE THIS STUDY (TO THE BEST OF MY KNOWLEDGE)

goto 63

HINTZMAN (19@@)

PICKS RANDOMLY GENERATED EXEMPLARS TO TRAIN HIS MODEL AND

ONLY PREDICTS THE BASIC TRENDS (GENERALIZE CONDITION

MORE ACCURATE THAN

CONTROL CONDITION) DESCRIBED IN THE PAPER

RECOGNITION MEMORY AND CLASSIFICATION

In testing models of category learning, we may examine how the classification of a given test pattern depends on ~~the~~ subjects' memory for ~~remembering~~ specific exemplars that were shown during training. Such "recognition memory" can be tested by asking subjects to judge whether a test pattern is an "old" training instance, or a "new" instance not experienced during training. Prototype theories, which assume that people extract only a mean centroid from the training exemplars, expect a strong correlation between classification and "old" judgments for test exemplars, since both decisions presumably could only be based on the distance of the exemplar from the prototype.

An experiment by Hayes-Roth and Hayes-Roth (1977) examined this issue: Over a variety of test patterns, they found a surprisingly low correlation between subjects' classifications and their Old (vs. New) judgments. It was of interest to see whether the configural-cue model could duplicate this surprising lack of correlation between classification and recognition memory. The training patterns used by Hayes-Roth and Hayes-Roth were descriptions of people who varied along three dimensions: age, marital status, and education level. Each of the three dimensions had four possible feature values. Assigning the numbers 1 through 4 to these values(~~at random~~ ~~[illegible]~~ done randomly for each subject), a training pattern may be represented by a three digit number. Thus, for a given subject '123' might refer to a stimulus pattern ("person description") who is '30 years old, married, with a college education.' Subjects were trained to classify such individuals as belonging to either Club 1, Club 2, or to neither Club. The assignment of individuals to clubs was determined by the following rules: a majority of 1's with no 4's *(e.g., 112,131)* signified membership in Club 1, whereas a majority of 2's with no 4's *(e.g., 212, 221)* indicated Club 2. An equal number of 1's and 2's with no 4's indicated membership in either club 50% of the ~~time~~ cases. If any ~~4's~~ 4-values were present, the person belonged to neither club. The ~~3's~~ 3-values were irrelevant features. Specific patterns were presented with widely varying frequencies. The most prototypical category members *(e.g., 111, 222, 333, 444)* were never presented during training, but were tested for subsequent recognition and classification. Training consisted of a single pass through a collection of @ patterns exemplifying these rules (see Hayes-Roth & Hayes-Roth, 1977, Table @, p. @, also Nosofsky, 1988, Table @., p. @).

Following training, Hayes-Roth and Hayes-Roth tested subjects' ability to recognize test patterns as being New or Old as well as to classify these same patterns as describing a

Club 1 or a Club 2 member. Their principal finding was that the accuracy of classifying a test item correlated only poorly (@r=@@) with its recognition. For instance, the non-presented category prototypes, *111* and *222*, were given the highest classification ratings; yet, these prototypes were rarely recognized as "old" training instances. Also, specific exemplars which were presented many times during the training session *(e.g., 112, 121)* received the highest recognition ratings, but weaker classification ratings.

Implications for Exemplar Storage Models

The low correlation between recognition memory and categorization was earlier thought to embarrass exemplar ~~storage~~ models (e.g., Anderson, Kline, & Beasley, 1979). The reasoning was that since classification in ~~such~~ exemplar models is based on memory for stored instances, whether an instance is classified accurately would seem intuitively to depend on whether it was remembered. However, Nosofsky (1988) showed that the exemplar-storage context model could account for the Hayes-Roth and Hayes-Roth data if recognition memory was assumed to be based on the summed similarity of a test pattern to *all* stored exemplars. italics This recognition rule was used previously in the SAM model of Gillund and Shiffrin (1984). Nosofsky's ~~amended~~ recognition ~~response~~ rule allowed the exemplar model to correctly predict~~ed~~ both the highly accurate classification of category prototypes, *111* and *222*, as well as the ~~high~~ accurate recognition of frequently-presented exemplars (see Table 6).

Insert Table 6 about here

We were interested to see whether the configural-cue network could also account for these data. The model was put through the training procedure of the Hayes-Roth and Hayes-Roth experiment. Patterns were presented in a random order to the network for one complete pass through each of the exemplars, with their appropriate frequencies. The network had four input nodes (cue values) for each of the three dimensions (12 total) plus 48 configural-cue nodes corresponding to the observed pair-wise combinations of the feature values. These input nodes were connected to three output nodes, representing the Club 1, 2, and Neither classifications. Reinforcement of the Club 1, Club 2, and Neither classifications occured by reinforcing the appropriate output node with $\lambda_j eq\, 1$ and the other two with $\lambda_j eq\, 0$. Exemplars in "either" the Club 1 or Club 2 categories were reinforced with either Club 1 or Club 2 membership with a 50% probability, corresponding to the Hayes-Roth and Hayes-Roth procedure.

The model's probability of assigning a test pattern to one of the two clubs was calculated from with the ratio of its activation to that of the other club node (the Neither node was not included because subjects were not given this option during testing). On the other hand, ~~the~~ ~~our~~ the model's recognition-memory rating for a test pattern was predicted from the summed activation of all three output nodes (including the Neither ~~node~~ category) to that pattern. This response rule is similar to Nosofsky's (1988) method for predicting recognition ratings with the context model.

During testing, subjects in the Hayes-Roth and Hayes-Roth experiment were required to classify each pattern as belonging to Club 1 or 2, and to rate on a 5 point scale their confidence in the accuracy of their classification. Unfolding this into a 10-point scale (with Club 1 negative) and normalizing the scores across all test patterns yielded the z-scores reported by Hayes-Roth and Hayes-Roth (see righthand columns of Table 6). The recognition scores were similarly computed: each test pattern was ~~tested~~ judged ~~recognized~~ as Old or New with confidence rated on a 4-point scale. Unfolded "Oldness" z-scores across test items reported by Hayes-Roth and Hayes-Roth are also given in Table 6. Figure 9 compares the

Insert Figure 9 about here

three models' predictions of recognition and classification ratings to test patterns with the observed ratings (transformed z-scores) given by subjects. Model predictions in Figure 9 are based on an average of multiple simulated training sequences, each with a different random sequential order of the training patterns. The simulations were done with the learning rate parameter, β, arbitrarily chosen to be 0.01; however, variations in this learning rate had ~~to~~ a negligible effect on the rank-order of the predictions.

While the single-cue model correctly predicted subject's classificatory responses with a rank order coefficient of 0.89, it predicted recognition memory with considerably less success (rank order coefficient of 0.82). In contrast, the configural-cue network model was more successful overall. Predictions of classificatory responses agreed with the data with a rank order correlation ~~=~~ of 0.90; importantly, the accuracy of its recognition predictions significantly improves over that of the single-cue model (rank order correlation = 0.94). These fits may be compared with those of Nosofsky (1988) who reported that the context model predicted the classification data with a Spearman rank-order correlation of .95 and predicted the recognition data with a correlation of .94 (see Table 6). [6] Nosofsky's model, however, requires at least one free parameter whereas the configural-cue network's predictions are largely parameter free (variations in β having a negligible effect over a wide range of values).

We also compared the predictions of the configural-cue model to the predictions of the feature-pattern model. The feature-pattern model's predictions for both sets of data were very similar to those for the configural cue model, and yielded no discriminating comparisons.

While the several models reveal some shortcoming in fitting these data, we will not analyze them further since there is reason to doubt the reliability of some aspects. Particularly troubling in the data of Table 6 are the classificatory z-scores which seem to cluster into three regions, those for Club 1 patterns (below -2.2), for Club 2 patterns (around 2.0), and for ambiguous patterns (around 0). The tight clustering of scores around these means suggest that subjects were not using the full 10-point classificatory scale discriminatively, but essentially were collapsing it into three points, reflecting judgments of Club 1, Club 2, or "Can't tell." Another reason to treat the z scores of Table 6 with caution is ~~Note also~~ that when response probabilites near .98 or (.02) are transformed to normal deviates (z scores), variations in response probability that are customarily considered trivial (say, from .98 to .99) produce large-appearing differences in normal-deviate scores. Many of the differences in classificatory z-scores among the patterns within the three clusters in Table 6 seem to be of this inconsequential variety. Therefore, we will not attempt to fit the models to any further break-downs of these data.

While the data are not all that we might hope for, the main conclusion of our theoretical exercise nonetheless remains. The configural-cue network model with no free parameters does a creditable job of fitting the Hayes-Roth and Hayes-Roth data on classification and recognition memory; importantly, it implies the surprisingly low correlation of these two measures.

BASIC LEVELS IN HIERARCHICAL CLASSIFICATION LEARNING

Basic Levels in Category Hierarchies

Among natural categories that stand in hierarchical relations to one another, Rosch and colleagues (Rosch, X & Y, 1976; Rosch & Mervis, 1978) found that categories at an intermediate level of abstraction are usually psychologically preferred or primary. They referred to this preferred level as the "basic level." Examples would be the term *bird* in the hierarchy *animal, bird, robin* or the term *apple* in the hierarchy *fruit, apple, mackintosh apple*. For natural categories, several behavioral indicators are correlated with the basic level. The basic level term is usually learned earlier by children, used more often in spontaneous labeling of an object, and verified more rapidly against a picture of the object. Basic categories are alleged to be the highest level for which one can form a recognizable composite image, because they often share perceptible parts. It is also the highest level for which standard interactions with objects share common actions, e.g., sitting on a *chair*. When asked to list common attributes shared by members of a class, people list relatively few attributes for superordinate categories (like *vehicles, furniture*); many more attributes for basic categories (*car, chair*), ~~and~~ but only a few more for subordinate categories (*coupe, lounger chair*).

Basic level categories appear to be special because they capture many regularities or patterns of correlations among the attributes of exemplars of the categories. A system of categories has two forces acting on it: on the one hand, one wants to have only a few categories so that some cognitive economy is achieved; on the other hand, one wants informative categories, meaning that by knowing that an object belongs to a given category one also knows many of its other attributes. In other words, one wants relatively few categories that are informationally rich, which maximize the feature similarity of objects in the category and minimize the similarity of objects in different classes. These desiderata are incompatible and categories ~~T~~trade-off how well they satisfy both goals. The basic level appears to be that level of generality which attains ~~in~~ a satisfactory compromise of both goals.

For present purposes, the relevant aspect of basic level phenomena is that many are well explained by our associative network model. Elsewhere (Gluck, Corter, & Bower, 198x; Corter, Gluck, & Bower, submitted) we have detailed the model's analysis and explanation of basic level phenomena. The key idea that mediates the basic level advantage stems from the LMS learning rule (cf. Eq. 1) which is sensitive to both cue validity and category validity. Cue validity is the extent to which presence of a given cue or feature predicts presence of a category (e.g., *having feathers* predicts *bird*). This association is directly incremented by Eq. 1 whenever a cue and category co-occur. Category validity is the converse -- the extent to which presence of a category predicts presence of the feature. This aspect is tracked by the cue-competition component of Equation 1 ~~that~~ reflected in overshadowing: if several features are already strongly predicting a given category (due to their higher validities or frequencies), they will crowd out or drive down the associative strength of some other cue that co-occurs with them and that also predicts the category.

The network model has been applied to laboratory experiments involving learning of artificial hierarchies of categories. An experiment by Hoffman and Zeissler (1983) may be used as illustration. They trained subjects to classify line drawings of rocket space ships that differed in their overall shape, bottom edge, and type of porthole (see Figure x). Three groups of subjects were taught to classify these eight pictures according to one of three category hierarchies shown in Figure x. The feature-category associations were designed so that the expected easiest level to learn would be the top, middle, and bottom in Hierarchies I, II, and III, respectively. The optimal level category was the one for which all instances of the category shared values on one or more feature dimension. For example, in Hierarchy I the top level is expected to be most easily learned because instances of the top-level categories share the same overall shape (angular vs. rounded). In Hierarchy II the middle level was expected to be optimal because instances in those categories share two features (porthole and shape) which contrast with the neighboring categories at that level. In Hierarchy III bottom level categories should be learned fastest because only three have a recognizable shape. ?

Insert Figure x about here

In Hoffman and Zeissler's experiment, subjets ~~saw~~ were presented with single stimuli out of the array and were told at what level of the hierarchy to classify ~~each~~ each. Response accuracy and latency were recorded. Subjects were trained to criterion, concurrently at all three levels.

The authors primarily reported their results in terms of errors during learning and response latencies over the final six blocks of trials at each level. The major results were clear in showing that the optimal classification level (with fewest errors and fastest response times) was the top level in Hierarchy I, the middle level in Hierarchy II, and the bottom level in Hierarchy III.

The complete ordering of levels according to average classification latency was: top-middle-bottom in Hierarchy I; middle-bottom-top in Hierarchy II; and bottom-middle-top in Hierarchy III. It is of interest to se~~x~~e whether the configural cue model can ~~has the wherewithal~~ ~~to~~ account for these complex set of results.

In applying the configural model to such results, we used 10 single-cue input nodes (corresponding to the 4 bottom edges, 4 porthole types, and 2 overall shapes) and the 20 pairwise configural cues that occurred in the 8 training stimuli. These 30 input nodes were fully connected to 14 output nodes representing the category responses (2, 4, and 8 at the top, middle, and bottom levels). the model was then applied to a simulation of the Hoffman-Zeissler Training procedure (for details, see Corter et al., submitted).

The simulated learning curves are shown in Figure y for a single learning rate (*BETA* = .025)). The trends of the theoretical curves closely follow the observations reported by Hoffman and Zeissler. In particular, the percentage correct responses over the last two blocks of trials are .96, .93, and .88 for the top, middle, and bottom levels in Hierarchy I, respectively; .84, .93, and .88 for top, middle, and bottom levels of Hierarchy II; and .64, .69, and .83 for the top, middle, and bottom levels of Hierarchy III. Assuming that correct naming speed is directly related to percentage correct responses, the predicted and observed orderings correspond perfectly across the three levels for each hierarchy. It is important to note that these relationships among associative strengths at different levels in different hierarchies are practically independent of the value of the learning rate, *BETA*. They are implied by the way the structure of the feature-to-category contingencies are dealt with

~~mapped into~~ by the cue competition model.

Insert Figure y about here

Elsewhere (Corter et al., submitted) we successfully applied the configural cue model to three other experiments studying the learning of three-level hierarchies of nested artificial categories--one by Murphy and Smith (1982) and two new experiments by Corter et al. (submitted). In each case feature-category contingencies similar to those in Hierarchy II (see Figure x) of Hoffman and Zeissler (1983) were investigated. As expected, the middle-level categories proved to be learned most quickly and ~~processed more easily~~ classified most quickly. The middle level is learned more rapidly in Hierarchy II because categories at that level are characterized by presence of a single feature (type of porthole); on the other hand, the superordinate level in Hierarchy II is defined by a disjunction of basic features (e.g., a *ril* has either a square or round porthole). The reader may wonder whether this structure characterizes hierarchies of natural categories for which Rosch et al. (1976) made their observations about basic level categories. Indeed, a number of investigators (Rosch et al., 1976; Smith & Medin, 1981); Corter et al., submitted) have observed this aspect of the typical hierarchies (e.g., *furniture, clothing, fruit, animal*). The superordinate category appears to be a disjunction of basic-level categories that have little in common except for nonperceptible, functional properties. Of course, this disjunctive character is why people cannot list many common attributes of superordinate categories nor image or act upon a common form.

Let us briefly, retrace our argument here. We have shown that when hierarchies of categories are structured as in Hierarchy II, the middle level is easier to learn by people and the network model. It is further argued that natural hierarchies that have preferred categories at the middle level are ~~visually~~ usually structured like Hierarchy II. Therefore, we suggest that if people learned categories as the network model expects, then the middle level would be learned quickest, and responded to most accurately and quickly. In turn that would explain a number of the behavioral observations offered in support of the basic level. Once a given level in a hierarchy is easy to learn, then other sequelae follow: it would be used more frequently by adults so that it would be acquired earlier by children and its label would tend to be shortened. The advantage of this structural-learning hypothesis is that it can accommodate to deviations from a simple basic level hypothesis caused by special expertise, training, and cultural use.

[End of section]

HYPOTHESIS TESTING IN RULE INDUCTION

MISCELLANEOU COMMENTS ON ADDITIONS NEEDED FOR SHJ

Shepard Quote re this data: "Models of conditioning of cues are not alone sufficient but in addition something like the abstraction and formulation of rules is apparently involved."

WHAT TO DO DIFFERENTLY FROM JML PAPER: As this is the last section of the paper (before discussion), I think that we can use these data to point out TWO LIMITATIONS to (suggestive of future work):

* 1. LIMITS OF SING + PAIRWISE CODING

Start first with application of same model used in other sections (singlets + doublets): Note produces correct order but does not learn type VI which requires triplets

ALSO: Types IV and V are not learned perfectly with only sing + doublets either. WHY?

Just because a task is Linearly Separably does not mean that the LMS network will reduce MSE to 0. LS implies that a THRESHOLD rule will solve perfectly. We have adopted a direct (raw) RATIO RESPONSE RULE. But note that exponential/logistic transform (ala Gluck - Bower 1988) would take care of this.

So we add Triplets for COMPLETE POWER SET (singlets+doub+triplets)

With RATIO rule learns all -- but 2 vs. 3 misorders..

* 2. NO ENCODING OF DIMENSIONAL STRUCTURE

----

* FOCUS ON II vs III -- do not ignore, is important

suggests network lacks "dimensionality sensitivity"

-- DESCRIBE DIMENSIONALITY ISSUE -- vs. INTRA/EXTRA Shifts

THUS: Model ignores dimensions and can not account for learning about dimensions that is independent of S-R associations.

--- CURRENT MODEL: Subsitutive dims represented ONLY

as sets of additive features

syntax error file -, between lines 3116 and 3247 Include not yet implemented file -, between lines 3116 and 3411

NotethatanalysesofothertasksinpapershowthatTripletsorQuadshavelittleeffectonpredictions..FROMHER

se

Examplesofthesixtypesofclassificationsusedby Shepardetal. (1961).*From ''Learning* and*memorizationofcla*

sh

Meansquarederror $\frac{plotted}{trials}$ *oflearningthesixclassificationtypesofShepard ,Hovland ,&Jenkins* (1961)*usinga*

simulations

to Shepard's Theory of Stimulus

Generalizationthissectionpulled (*as is*) 1991.*ppShepard* (1987)*describesarationalmotivation* for*theexponen*

Gluck,

have

*regions.Followingasingletrainingtrial ,theindividualispresumednoknowledgeofthe*location*oftherelevantconsequ*

Regions and Configural-

understanding

cues.ppShepard'sapproachcanalsobe applied howtheconfigural−cuemodelsolvescategorizationproblems.

need transition (LS/NLS cut from psych sci article)

Shepard's 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 & 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 which might extend the range of phenomena deducable by either theory alone. We briefly note three such possibilites here. First, Shepard 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 ~~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 multi-layer networks. In spite of their complexity, multi-layer 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 multi-layer network.

Comparison to Single-Cue Network Model

INSERT HERE?? ANALS OF CONFIGCUE APPLIED TO GB88

Limitations of Model

Dimensional Salience

Lawrence/Nonreversals, intra/extra, SHJ, II/III

Feature Extraction

wheredothefeaturescomefrom?.lh Configural Explosion with Many Cues but singlets + doublets almost always suffice

future
stillausefultool for*poiting way theorydevelopment* exps

Other

GORDON: OTHER LIMITS TO NOTE...

Concluding Remarks

THIS IS TAKEN FROM OUR CONFIG-CUE COG SCI PAPER

By expanding the representation of stimuli to include pair-wise configurations of features, the network model appears to account for a wider range of learning results from both the animal and human learning literatures. Some of this success can be traced to its using a similarity metric like that of Medin & Shaffer, viz., an implicit exponential decay relationship between stimulus similarity and psychological distance (number of feature mismatches). The configural-cue model has several obvious limitations, including the exponential growth of input nodes with increasing pattern size. Nevertheless, we believe that this model is interesting for four reasons. First, it is simple, understandable, and accounts for a surprisingly wide range of empirical phenomena. Second, it is theoretically parsimonious and uses assumptions for which independent evidence already exists. Third, its successes are instructive in identifying empirical phenomena which can be explained as emergent from the same elementary, associative processes found in lower species. Fourth, explanations of the failures of this model can suggest more sophisticated versions of the network model. Such failures may also indicate performances arising from an entirely different class of learning mechanisms, i.e., the rule-based or symbolic processes which have been well studied by cognitive psychologists.

Footnotes

Correspondence regarding this article should be addressed to: Dr. Mark A. Gluck, Center for Molecular & Behavioral Neuroscience, Rutgers University, 197 University Ave., Newark, New Jersey 07102. Email: *gluck@pavlov.rutgers.edu* For their thoughtful comments and advice on this work we are indebted to: James Corter, W. K. Estes, Robert Nosofsky, Jay McClelland, Doug Medin, Misha Pavel, Robert Rescorla, and Allan Wagner. This research was supported by a Henry Rutgers Fellowship to MAG, an Office of Naval Research grant (@@) to MAG,

[1]

[2] One could imagine a short-term memory process whereby the weights change a large amount so as to immediately reduce the discrepancy to zero, thus guaranteeing a correct response were the same pattern to be presented again immediately. In this approach, one would have to assume further that with the passage of time and other interfering patterns, the new weights in long-term memory would decline to approximately those given by Equation 3 with some smaller value of β.

[3] A qualification to the configural hypothesis is that some brief time is needed for the configural units <AB> to be recruited after presentation of the AB compound. For example, Kehoe and X (199x) found that rabbits could learn a negative-patterning eyeblink discrimination (tone+, light+, tonelight-) only if the CS-UCS interval was longer than 400 msec; with a shorter 300 ms interval, animals apparently triggered their conditioned eyeblink to any signal first sensed on a trial and never learned to wait for both cues in order to inhibit

response to the compound. A similar minimum interval would doubtless be needed to train adult humans to respond to either of two signals but not to both signals presented sequentially in order to avoid an aversive outcome. The existence of this configural recruitment-time will be ignored in our applications, since in these experiments the human subjects were permitted to respond at their own liesurely pace after fully inspecting each stimulus pattern.

4. @ note alternative: inside learning eq. or outside , cf. g&b1990 @ --> all simulations were done with the ratio rule outside learning

5. We note that this prediction hinges crucially on the specific set of stimuli constructed by Medin and Schwanenflugel to realize the LS and NLS classification tasks. More generally, across a range of problems, the relative difficulties of linearly separable and non-linearly separable tasks should vary; in terms of ecological statistics, most linearly separable tasks would probably be expected to be easier than non-linearly separable tasks.

6. These rank-order correlations reported by Nosofsky (1988, p. @) are for the single-parameter version of the context model, with the similarity parameter, s, held constant across classification and recognition. The values given in Table 6 and Table @ of Nosofsky (1988) are for the 2-parameter version which yielded a rank order correlation of .97. (Robert Nosofsky, *personal communication*, August 22, 1990).

References

Table 1 t1.

///_/_/a/b

/c /d.TE.bp.TSbox ,center ,tab (/);cfBsssssssssscfBssssssssscssssssssssscssss | cssssccsss | ccssscccc | ccccccccc | cccccccc

/!n2!/1/0/0/0/.94/.93/Transfer /!n 3!/1/1/1/1/.50/.57/

Patterns

FIND MISSING TRANSFER MODEL PREDICITONS..