Journal of Experimental Psychology:
Learning, Memory, and Cognition
1992, Vol. 18, No. 2, 211-233

0278-7393/92/$3.00

Combining Exemplar-Based Category Representations and Connectionist Learning Rules

Robert M. Nosofsky, John K. Kruschke, and Stephen C. McKinley
Indiana University

Adaptive network and exemplar-similarity models were compared on their ability to predict category learning and transfer data. An exemplar-based network (Kruschke, 1990a, 1990b, 1992) that combines key aspects of both modeling approaches was also tested. The exemplar-based network incorporates an exemplar-based category representation in which exemplars become associated to categories through the same error-driven, interactive learning rules that are assumed in standard adaptive networks. Experiment 1, which partially replicated and extended the probabilistic classification learning paradigm of Gluck and Bower (1988a), demonstrated the importance of an error-driven learning rule. Experiment 2, which extended the classification learning paradigm of Medin and Schaffer (1978) that discriminated between exemplar and prototype models, demonstrated the importance of an exemplar-based category representation. Only the exemplar-based network accounted for all the major qualitative phenomena; it also achieved good quantitative predictions of the learning and transfer data in both experiments.

One of the major current models for explaining performance in arbitrary category learning paradigms is the context model proposed by Medin and Schaffer (1978) and elaborated by Estes (1986a) and Nosofsky (1984, 1986). According to the context model, people represent categories by storing individual exemplars in memory and make classification decisions on the basis of similarity comparisons with the stored exemplars. The context model has proved to be successful at predicting quantitative details of classification performance in a wide variety of experimental settings and has compared favorably with a variety of alternative models, including prototype, independent-feature, and certain logical-rule based models (see Medin & Florian, in press, and Nosofsky, in press-a, in press-b, for reviews).

However, some shortcomings of the context model were recently demonstrated in series of probabilistic classification learning experiments conducted by Gluck and Bower (1988a) and Estes, Campbell, Hatsopoulos, and Hurwitz (1989). In these studies, an adaptive network model was shown to provide superior predictions of classification learning and probability judgment to those of the context model.

This work was supported by Grants BNS 87-19938 from the National Science Foundation and PHS R01 MH48494-01 from the National Institute of Mental Health to Robert M. Nosofsky, and by BRSG Grant RR 7031-25 from the Biomedical Research Support Grant Program, Division of Research Resources, National Institutes of Health to John K. Kruschke.

We thank Jerry Busemeyer, William Estes, Mark Gluck, David Shanks, and an anonymous reviewer for helpful criticisms and suggestions. We also thank Larry Barsalou for his criticisms and encouragement.

Correspondence concerning this article should be addressed to Robert M. Nosofsky, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent to nosofsky@ucs.indiana.edu.

The purpose of our research was to follow up on the Gluck and Bower (1988a) and Estes et al. (1989) studies and continue this line of investigation by comparing adaptive network models and exemplar models of classification learning.

An important new direction was to test an integrated, exemplar-based network model, which combines key components of both modeling approaches (Hurwitz, 1990; Kruschke, 1990a, 1990b, 1992; for related ideas, see Estes, 1988). The Medin and Schaffer (1978) context model and the Gluck and Bower (1988a) adaptive network model can be contrasted on two major dimensions. The first concerns the nature of the basic units that are stored in memory. According to the context model, people store individual exemplars in memory, and classification decisions are based on the similarity of a probe to the stored exemplars. By contrast, in the Gluck and Bower (1988a) adaptive network model, people learn associations between individual features and the alternative categories. A second dimension of contrast concerns the nature of the learning rule. In the Gluck and Bower (1988a) adaptive network model, an error-driven, interactive learning rule is assumed. By contrast, learning in the context model is not error-driven but consists simply of the gradual accumulation of exemplars in memory. The exemplar-based network model combines what we believe are key advantages of each approach, namely, an exemplar-based category representation that incorporates an error-driven learning rule.

We organize this article by first reviewing the main experimental paradigms and models tested by Gluck and Bower (1988a) and Estes et al. (1989). After suggesting some limitations of the particular versions of the context model and the adaptive network models that were tested, we discuss various elaborated versions of these models, including the exemplar-based network model. These models are then compared in two experiments that extend the original Gluck and Bower (1988a) and Estes et al. (1989) studies. Experiment 1 demonstrates the advantages of assuming an interactive, error-

driven learning rule, whereas Experiment 2 focuses on the advantages of an exemplar-based category representation.

Review of the Experiments

In the Gluck and Bower (1988a) and Estes et al. (1989) experiments, subjects learned to classify stimuli varying along four binary-valued dimensions into two categories. The categories were defined by independent probability distributions over the set of dimension values composing the stimuli. The four positive values on Dimensions 1 through 4 are denoted by f_1, f_2, f_3, and f_4, respectively; the four negative or alternative values are denoted by f_1^*, f_2^*, f_3^*, f_4^*. (In some paradigms, the positive values represent the presence of particular features, and the negative values represent their absence, whereas in other paradigms the terms *positive* and *negative* are arbitrary.) The probability of each of the four positive dimension values, given the alternative categories, was as follows:

Feature	Category A	Category B
f_1	.6	.2
f_2	.4	.3
f_3	.3	.4
f_4	.2	.6

The probability of each of the four negative feature values was the additive complement of the positive feature probabilities, for example, $P(f_1^* \mid A) = 1 - P(f_1 \mid A) = .4$. Thus, for example, the probability of the stimulus composed of dimension values f_1, f_2^*, f_3, and f_4^*, given a Category A trial, was (.6) (1 − .4) (.3) (1 − .2) = .0864. Note that the base-rate probability of the categories was also varied. The prior probability of Category A was .25, and of Category B was .75.

On each trial, a category was sampled in accord with the base-rate probabilities, and a stimulus was then generated in accord with the category-feature probabilities listed earlier. Subjects judged whether the stimulus belonged to Category A or B, and feedback was then provided. In both studies a series of test trials was also used in which subjects were presented with single features of the training patterns and were asked to make either probability judgments or classification choices on the basis of the presence of these single features.

A key result of the Gluck and Bower (1988a) studies, which was by and large replicated by Estes et al. (1989), was that subjects exhibited a tendency toward base-rate neglect when presented with the single features on the test trials. Specifically, the experimental design ensured that the normative probability of Category A given feature f_1 was equal to the normative probability of Category B given feature f_1, namely, .50. However, subjects in both experiments estimated the probability of Category A given f_1 to be substantially greater than .50, apparently not taking full account of the differential category base rates. As we discuss later, these results were consistent with the predictions of the adaptive network models but contradicted the predictions of the context model.

Estes et al. (1989) extended Gluck and Bower's (1988a) findings in two major ways. First, whereas Gluck and Bower (1988a) demonstrated the base-rate neglect phenomenon by collecting direct probability judgments of categories given individual features, Estes et al. showed that the phenomenon also existed when classification choice data were collected. Second, Estes et al. (1989) used a fixed sequence of 240 item presentations during the classification learning phase that was common to all subjects. This experimental method allowed Estes et al. to fit the competing models to the sequence of trial-by-trial classification responses observed during learning. Estes et al. found that the adaptive network models provided superior quantitative fits to this sequence of classification learning data compared with those of the context model.

Review of the Models

In this section we briefly review the adaptive network model and the version of the context model that were tested in the previous studies of Gluck and Bower (1988a) and Estes et al. (1989).

Component-Cue Network Model

We refer to the adaptive network tested by Gluck and Bower (1988a) as the component-cue network model. As discussed by Gluck and Bower, the particular version of the component-cue network that is applied varies, depending on whether the dimension values or features are additive or substitutive in nature (Tversky, 1977). Additive features are those in which a particular feature is either present or absent, whereas for substitutive features, one of two positively existing values is present on each trial. Here we review the version of the component-cue model that is applied when substitutive features are used because we use stimuli with substitutive features in our experiments.

The component-cue model is illustrated in Figure 1. Each dimension m is coded by a pair of input nodes, which are activated according to the values on a presented pattern x. If the pattern has a positive value on dimension m, then the first input node in the pair is activated with a value of one and the second node activated with a value of zero, $a_{m1}(x) = 1$ and $a_{m2}(x) = 0$; whereas if the pattern has a negative value on dimension m, then the reverse activations occur. Following Estes et al. (1989), in cases in which there is a missing value on dimension m, such as occurs during the test trials when only single features are presented, both input nodes in the pair are set at zero.

In this study we consider an augmented version of the component-cue model that includes a *bias* or *context-coding* node at the input level. The activation of the bias node is set to one on all trials. As will be seen, the bias node is necessary if the component-cue model is to adequately characterize a certain base-rate sensitivity phenomenon that is observed in the transfer phase of our experiments. The bias node differs from the null node used by Estes et al. (1989), which was assumed to be activated only on those trials in which a null pattern (the stimulus with no features) was presented. (The null pattern is never presented during our training phases, thus no learning would occur on the null node.)

The activations on all input nodes are multiplied by the connection weights currently existing in the network, which are then summed to form outputs. The output received by category node A is given by

$$O_A(x) = \Sigma a_{mj}(x) w_{mj,A} + b_A, \quad (1)$$

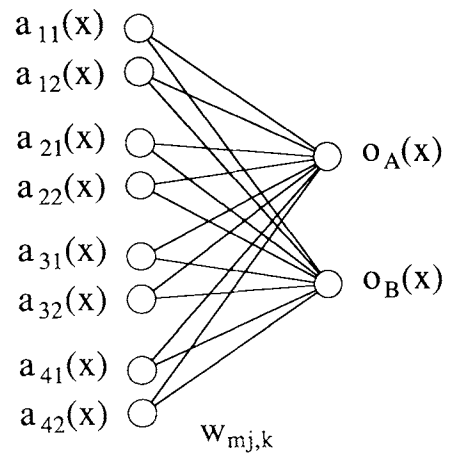

Figure 1. Illustration of Gluck and Bower's (1988a) component-cue model.

where $w_{mj,A}$ is the weight on the connection that links input node a_{mj} to category output node A, and b_A is the weight from the bias node to output node A. An analogous expression is used to compute the output received by category node B. The probability of a Category A response given presentation of pattern x is given by

$$P(A|x) = \exp[cO_A(x)]/\{\exp[cO_A(x)] + \exp[cO_B(x)]\}, \quad (2)$$

where c is a freely estimated scaling constant (Estes et al., 1989).

All weights in the network are initialized at zero. Learning of weights takes place through the delta or least-mean-squares rule (Widrow & Hoff, 1960). When Category A feedback is presented, output node A receives a teaching signal of $z_A = 1$, and output node B a teaching signal of $z_B = -1$, whereas when Category B feedback is presented the reverse teaching signals are provided. The error at output node A is given by

$$\delta_A = [z_A - O_A(x)], \quad (3)$$

and likewise for the error at output node B. All weights in the network are then updated by using the rule,

$$\Delta w_{mj,K} = \beta \delta_K a_{mj}(x), \quad (4)$$

where β is a learning rate parameter. The bias-node weights are updated by using an analogous formula, except they are allowed a separate learning rate parameter, β_b. Note that when $\beta_b = 0$, this augmented model reduces to the component-cue model tested by Gluck and Bower (1988a) and Estes et al. (1989).[1]

The component-cue model has the following three free parameters: learning rates β and β_b for updating the feature weights and bias-node weights, respectively; and the scale parameter c for transforming outputs to response probabilities.

Configural-Cue Model

Gluck and Bower (1988b) also proposed a configural-cue network model, in which the input nodes code not only the individual feature values that compose a stimulus but also all pairs of features, triples of features, and so forth. With respect to our Experiment 1 data, it turns out that the simple component-cue model outperforms the configural-cue model (see Appendix). By elaborating the configural-cue model with free parameters, we can find versions that basically reduce to the component-cue model but provide no additional explanatory power. For simplicity, therefore, we focus initially on the component-cue model and then treat the configural-cue model in more detail in Experiment 2.

Context Model

According to the version of the context model tested by Estes et al. (1989), on each trial a presented exemplar is stored in memory with some fixed probability. The probability that a pattern x is classified in Category A is found by summing the similarity of x to all previously stored exemplars of Category A and then dividing by the summed similarity of x to all previously stored exemplars of both Categories A and B, as follows:

$$P(A|x) = \sum_{a \in A} s(x, a) \Big/ \left[\sum_{a \in A} s(x, a) + \sum_{b \in B} s(x, b) \right], \quad (5)$$

where $s(x, a)$ denotes the similarity of x to exemplar a. Note that the learning process here is conceptualized in terms of the gradual accumulation of exemplars in memory.

The similarity of x to exemplar a is computed by using Medin and Schaffer's (1978) multiplicative rule:

$$s(x, a) = \prod_m s_m^{\delta_m(x, a)}, \quad (6)$$

where s_m $(0 \leq s_m \leq 1)$ is a freely estimated parameter reflecting the similarity of mismatching values on dimension m; and δ_m (x, a) is an indicator variable equal to one if x and a have mismatching values on dimension m, and equal to zero otherwise. (The δ_m in Equation 6 is not to be confused with δ_A of Equation 3.) In the version of the model fitted by Estes et al., all similarities s_m were set equal to a common free parameter s.

Nosofsky (1984, 1986) noted that Medin and Schaffer's (1978) multiplicative rule can be interpreted in terms of a multidimensional scaling approach to modeling similarity (Shepard, 1958, 1987). Let x_m and a_m denote the psychological values on dimension m of pattern x and exemplar a, respectively. The distance between those objects is computed by using the (weighted) Minkowski power model formula,

$$d(x, a) = \left[\Sigma \alpha_m |x_m - a_m|^r \right]^{1/r},$$

[1] Gluck and Bower (1988a) describe a model in which there is a single output node that receives a teaching signal of $z = +1$ when Category A feedback is provided and $z = -1$ when Category B feedback is provided. The version of the model that we describe with two output nodes, each receiving $+1/-1$ or $-1/+1$ teaching signals, is, under the present conditions, formally identical to the single-output node version used by Gluck and Bower (1988a).

where α_m is the *attention weight* given to dimension m. Common values of r in Equation 7 are $r = 1$, which yields the city-block metric; and $r = 2$, which yields the Euclidean metric (Garner, 1974; Shepard, 1964). This distance is converted to a similarity measure by using the transformation,

$$s(x, a) = \exp[-\kappa d(x, a)^p], \tag{8}$$

where κ is a general sensitivity parameter. Common values of p in Equation 8 are $p = 1$, which yields an exponential decay function; and $p = 2$, which yields a Gaussian decay function (Nosofsky, 1985; Shepard, 1958, 1987).

It is straightforward to verify that when $p = r$ in Equations 7 and 8, an interdimensional multiplicative similarity rule of the form used by Medin and Schaffer (1978) is yielded (e.g., see Nosofsky, 1986). This relation led Nosofsky (1984, 1986) to propose the *generalized context model* (GCM), in which exemplars are represented as points in a multidimensional space, with similarities among exemplars computed by using Equations 7 and 8. The critical idea we wish to emphasize here is that the differential similarity parameters that enter into the context model's multiplicative rule can be interpreted in terms of an attention process in which the dimensions are differentially weighted.

Comparison of the Models

As is apparent from the preceding review, numerous differences exist between the network and exemplar models. As stated in the introduction, one key difference involves their learning rules. In this section we describe the difference in learning rules in more detail, with emphasis on how it might underlie the models' differing abilities to account for sequential-learning and base-rate neglect phenomena.

Sequential-Learning Phenomena

In the network-model learning rule, the connection weights are adjusted proportionally to the error produced (Equation 4). As error is gradually reduced through learning, the weights change less and less on any given trial. By contrast, the version of the context model tested by Estes et al. (1989) used a learning mechanism in which the strength of an exemplar was incremented by a constant amount every time it was presented, regardless of the performance of the system. This difference between error-driven and constant-increment learning can result in different learning behavior for the two models. For example, suppose that each model has already learned that a particular input pattern is mapped to a particular category. When that pattern is presented again to the error-driven network model, the connection weights change very little because there is very little error. However, when the pattern is presented to the constant-increment context model, the memory strength for that pattern is incremented by the same amount as for all previous presentations.

The learning rule used in the network model also gave it a built-in sensitivity to recency, whereas the version of the context model tested by Estes et al. had none. In the network model, the weights that are operative on trial n will be influenced more by the stimulus pattern and feedback provided on trial $n - 1$ than by the same pattern and feedback provided 100 trials earlier (assuming equal errors in prediction at the two points in the learning sequence). The reason is that in the network model, multiple patterns activate common nodes, so newly presented patterns can modify and undo connection weights that were previously learned (cf. McCloskey & Cohen, 1989; Ratcliff, 1990). By contrast, in the version of the context model tested by Estes et al., all exemplars have the same influence regardless of their recency—one simply sums the similarity of a probe to all previously presented exemplars, with equal weight given to each exemplar. Assuming that recency effects occur in classification learning, this aspect of the context model is probably another reason for its shortcomings in predicting sequential learning data.

Finally, the version of the context model tested by Estes et al. assumed equal similarity parameters (or attention weights) for all stimulus dimensions, and the value of the weights was held fixed across the entire learning sequence. Because extensive previous tests of the context model suggest that differential weighting of the dimensions occurs (e.g., Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky, 1986, in press-b), and furthermore that similarities may change as a function of learning (e.g., Nosofsky, 1987), this assumption does not in general appear to be tenable. Nevertheless, as will be seen, differential attention-weighting does not appear to be critical for explaining sequential-learning phenomena in these experiments.

Base-Rate Neglect Phenomena

Consider the context model's prediction of the probability that, during the test trials, a subject classifies feature f_1 in Category A, $P(A \mid f_1)$. (Note that the object composed of only feature f_1 is logically distinct from the pattern $f_1f_2^*f_3^*f_4^*$.) Following Estes et al. (1989), we assume that similarity is computed only over those dimensions that have nonmissing values. Thus, the similarity of f_1 to an exemplar having feature f_1 is 1.0, and the similarity of f_1 to an exemplar having feature f_1^* is s. The context model predicts the following:

$$P(A \mid f_1) = [n_A + (N_A - n_A)s]/\{[n_A + (N_A - n_A)s] + [n_B + (N_B - n_B)s]\}, \tag{9}$$

where n_A and n_B are the relative frequencies of exemplars that have feature f_1 *and* are in Categories A and B, respectively; and N_A and N_B are the base rates of Categories A and B, respectively. [Note that $n_A = N_A P(f_1 \mid A)$ and $n_B = N_B P(f_1 \mid B)$.]

In the Gluck and Bower (1988a) design, $N_A = .25$, $N_B = .75$, $n_A = .25(.6)$, and $n_B = .75(.2)$. When the similarity parameter s in Equation 9 is zero, the context model predicts

$$\begin{aligned} P(A \mid f_1) &= n_A/[n_A + n_B] \\ &= .25(.6)/[.25(.6) + .75(.2)] \\ &= .50, \end{aligned} \tag{10}$$

which is the normative probability of Category A given f_1. This prediction was the one that Gluck and Bower (1988a)

ascribed to exemplar models and that was contradicted by their data (because subjects' probability judgments of Category A given f_1 were substantially greater than .50).

When the similarity parameter s in Equation 9 is nonzero, matters become even worse for the context model. In the extreme, when $s = 1$, the context model predicts the following:

$$P(A \mid f_1) = N_A/(N_A + N_B) = .25/(.25 + .75) = .25, \quad (11)$$

which is the base-rate probability of Category A. Values of s intermediate between 0 and 1 lead to predictions of $P(A \mid f_1)$ that are intermediate between .50 and .25, in contradiction to the observed data.

In sum, the context model assumes that people store individual exemplars in memory and compute summed similarities of patterns to these stored exemplars. This combination of assumptions leads the context model to predict that base rates will be used when subjects make probability judgments of categories when given individual features.

As explained by Gluck and Bower (1988a), the network model is able to predict the base-rate neglect phenomenon because of its interactive learning rule for adapting the connection weights. There is a sense in which the individual cues of the stimuli compete with one another to become associated with the alternative categories. Cues that are relatively better predictors become more highly associated with the category, that is, larger connection weights develop between those cues and the category. Because feature f_1 is a relatively poor predictor of the high-probability Category B, the connection weight from f_1 to Category A ends up being larger than the connection weight from f_1 to Category B, and the model therefore predicts the base-rate neglect phenomenon.

A Sequence-Sensitive Version of the Context Model

Most previous successful applications of the context model have occurred in classification *transfer* situations, in which performance is tested following the completion of classification learning. The results of Estes et al. (1989), however, suggest the need to augment the model with respect to its predictions of classification *learning*. One goal of our research was to develop and test some elaborated versions of the context model that might allow it to more accurately characterize processes of classification learning and trial-by-trial changes in category representations as a function of experience.

A reasonable starting proposal is that rather than giving equal weight to all exemplars in computing summed similarities, more recently presented exemplars ought to receive greater weight. Recency effects are ubiquitous in the memory literature, and such effects ought to be formalized in any model of the classification learning process. In previous tests of the context model the role of recency was not incorporated because the goal was to predict classification transfer data, and the precise sequence of training exemplars was randomized over subjects. However, in a fixed-sequence learning design such as the one used by Estes et al., recency effects may be of critical importance.

The recency-sensitive model assumes that

$$P(A \mid x) = \sum_{a \in A} M_a s(x, a) \Big/ \left[\sum_{a \in A} M_a s(x, a) + \sum_{b \in B} M_b s(x, b) \right], \quad (12)$$

where M_a is the *memory strength* associated with exemplar a. We assume that exemplar memory strength is an exponential decay function of lag of presentation,

$$M_a = \exp(-T lag), \quad (13)$$

where *lag* is the number of intervening trials between the presentations of pattern x and exemplar a, and T is a freely estimated time-rate decay parameter. Although not made explicit in the notation in Equation 12, we are treating presentations of the same exemplar on different trials as distinct memory traces, with each trace having its own memory strength. Thus, the summations are over all previous presentations of each exemplar a. (For evidence that more complex memory-weightings of each exemplar may be involved, see Busemeyer & Myung, 1988, and Myung & Busemeyer, in press).

A second elaboration of the context model involves an attempt to characterize processes that occur very early in the learning sequence. Early in the sequence, averaged classification probabilities for each trial tend to hover around .50 and only gradually move away from this starting point. The network model predicts such behavior because the connection weights are initialized at zero and are adjusted by small amounts as learning progresses. The standard context model, however, tends to predict classification probabilities early in the sequence that are too extreme. For example, suppose that on Trial 1 an exemplar from Category A is presented. Then the model would predict that the item presented on Trial 2 would be classified in Category A with probability 1.0 because it has some positive summed similarity to the exemplars of Category A, and zero summed similarity to the exemplars of Category B.

To address this shortcoming, we propose that there is some background noise in subjects' memory representations, and only after sufficient training do the summed similarities to stored exemplars overcome this noise. We formalize this idea by assuming the following:

$$P(A \mid x) = \left[\sum_{a \in A} M_a s(x, a) + B \right] \Big/ \left[\sum_{a \in A} M_a s(x, a) + \sum_{b \in B} M_b s(x, b) + 2B \right],$$

where B is a background-noise constant. Early in the learning sequence the background noise will dominate and classification probabilities will hover around .50, whereas later in the learning sequence the summed similarities will dominate and classification will be based on the experienced exemplars. The background-noise constant will play a major role even late in the learning sequence if similarity is low and a particular exemplar has been presented infrequently.

The sequence-sensitive context model (Equations 6, 13, and 14) has the following three free parameters: the similarity parameter (s), the decay rate (T), and the background-noise constant (B).

The sequence-sensitive version of the model has little to say about the base-rate neglect phenomenon demonstrated by Gluck and Bower (1988a) and Estes et al. (1989). We must point out the very simple possibility that these results could be reflecting a response bias. Inspection of the test-trials data reported by Gluck and Bower (1988a), for example, reveals that the observed probability judgments lie almost uniformly above the normative ones predicted by the exemplar model. With the addition of a response-bias parameter, the exemplar model could fit these data quite well. Furthermore, there are independent grounds for expecting that a response bias might operate during the test trials. The probability of Category A given *any* of the individual features never exceeds the probability of Category B. Thus, without bias operating, subjects would never judge A to be more likely than B during the test trials. Because work by Parducci (1974) suggests that subjects often shift their response criteria in an attempt to equalize their use of alternative category labels, the possibility that some form of response bias was operating in the Gluck and Bower (1988a) and Estes et al. (1989) experiments should be examined. We test this possibility by collecting a more detailed set of test-trials data in our study.

An Exemplar-Based Network Model

A shortcoming of the sequence-sensitive context model proposed in the previous section is that the model still lacks any form of error-driven, interactive learning. Furthermore, there is no mechanism for how the dimensional attention weights are updated trial by trial. Kruschke (1990a, 1992) recently proposed an integrated model in which key components of the GCM are implemented within a multilayered connectionist network (for closely related work, see Hurwitz, 1990). This integrated model overcomes the shortcomings of the context model noted earlier and, we will argue, offers significant advantages over the adaptive network models tested by Gluck and Bower (1988a) and Estes et al. (1989).

Kruschke's (1990a, 1992) model is known as ALCOVE, which stands for *attentional learning covering map*. The ALCOVE model, illustrated in Figure 2, consists of (a) a set of input nodes that code the values on the dimensions composing a given input pattern, (b) a set of hidden nodes that code locations in the multidimensional space in which the exemplars are embedded, and (c) a set of category output nodes that code the degree to which the alternative categories are activated.

Each hidden node is activated according to its similarity to a given input pattern, in which similarity is computed as in the GCM. For example, if a given hidden node codes location j in the multidimensional space, then when the network is presented with pattern x, hidden node j's activation is equal to the similarity between locations x and j in the multidimensional space. In a covering map version of the model, numerous hidden nodes are randomly scattered throughout the entire multidimensional space, whereas in a pure exemplar-based version, which we focus on here, hidden nodes are established at only those locations where individual exemplars were presented during training. For clarity, we hereinafter refer to the hidden nodes in the exemplar-based version of ALCOVE as *exemplar nodes*.

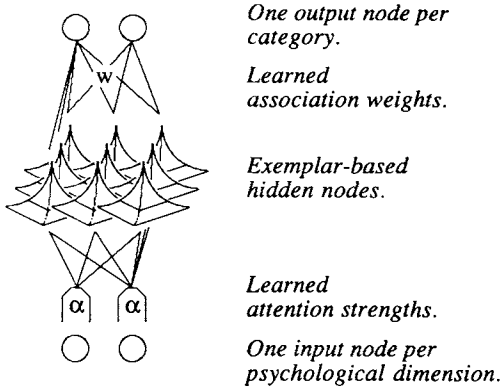

Figure 2. Illustration of an exemplar-based version of Kruschke's (1992) ALCOVE model. (Input nodes code values on the respective stimulus dimensions, and these values are gated by the attention weights, α_m. Hidden nodes code locations in the multidimensional space in which the exemplars are embedded; each hidden exemplar-node is activated according to its similarity to the input pattern. The pyramids show the activation profile of a hidden node in which there is a city-block metric with exponential decay [$r = p = 1$ in Equation 15]. The hidden exemplar-nodes are connected to category output nodes by the association weights, w_{jK}. Learning of the attention weights and the association weights takes place by using the generalized delta rule.)

As illustrated in Figure 2, the input nodes are gated by dimensional attention weights. These attention weights enter into the activation function that computes the similarity between an input pattern and each exemplar node. Suppose, for example, that pattern $x = (x_1, x_2, \ldots, x_M)$ is presented. Then each exemplar-node j is activated by using the function,

$$a_j(x) = \exp\{-\kappa[\sum \alpha_m |x_m - h_m|^r]^{p/r}\}, \qquad (15)$$

where x_m is the value of the input pattern on dimension m; h_m is the value of (hidden) exemplar-node j on dimension m; α_m is the attention weight on all connections that link input node m to the various exemplar nodes; κ is a general sensitivity parameter; and r and p reflect the distance metric and similarity function that determine activations in the multidimensional space. In this article we assume a city-block distance metric ($r = 1$) and an exponential decay similarity function ($p = 1$; Shepard, 1987).

The function for computing the exemplar-node activations (Equation 15) is formally identical to the similarity rule assumed in the generalized context model (GCM; see Equations 7 and 8). This exemplar-based version of ALCOVE, however, departs from the GCM in two major respects. First, instead of simply summing the similarity of a pattern to the exemplars of the alternative categories, ALCOVE assumes that *associations* are learned between the exemplar nodes and the categories. These associations, which are allowed to be positive or negative, are modeled by the association weights W_{jK} that link each exemplar node j to each category output

node K (see Figure 2). The output to a category node is then given by

$$O_K(x) = \sum a_j(x)W_{jK}, \quad (16)$$

where the sum is taken over all exemplar nodes. The association weights are learned by this exemplar-based network by using the same error-driven, interactive principles that are incorporated in the adaptive network models of Gluck and Bower (1988a) and Estes et al. (1989).

A second key contribution of ALCOVE is that it provides a mechanism for how the attention weights are learned. Although the attention weights have been critical in allowing the GCM to achieve its precise quantitative fits to classification transfer data, no learning mechanism was proposed. Thus, ALCOVE provides a model of attention-weight learning, whereas the GCM has none.

Predictions of classification response probabilities are generated in ALCOVE in the same manner as described previously for Gluck and Bower's (1988a) component-cue model (see Equation 2). Trial-by-trial learning of the association weights and the attention weights in the model takes place by using the generalized delta rule (Rumelhart, Hinton, & Williams, 1986). The details of the learning algorithm are reported by Kruschke (1990a, 1992).

The version of ALCOVE described in this section has the following four free parameters: an overall sensitivity parameter (κ) that enters into the similarity function for computing the exemplar-node activations (Equation 15); a scale parameter (c) for transforming category outputs into classification response probabilities (Equation 2); and learning rates (β and β_{ATT}) for adjusting the association and attention weights, respectively. In applications in our study, however, the attention-weight learning rate is set at zero to equalize the number of free parameters used by ALCOVE and the competing models.[2]

Because the exemplar-based network model has mechanisms for learning the association weights, it may prove capable of accurately fitting trial-by-trial data obtained in fixed-sequence learning designs. Furthermore, because it incorporates principles of error-driven, interactive learning, it may accurately predict the tendency toward base-rate neglect that is observed when subjects are tested with individual features of the training patterns. These questions are pursued in Experiment 1 of our study. Unlike the component-cue adaptive network, ALCOVE incorporates an exemplar-based category representation. We demonstrate the advantages of assuming such a representation in Experiment 2 of our study.

Experiment 1

The purpose of Experiment 1 was to test the component-cue network model, the sequence-sensitive context model, and the exemplar-based network model on their ability to quantitatively predict learning and transfer performance in a fixed-sequence, probabilistic classification paradigm. The experiment involved partial replications and extensions of the experiments reported previously by Gluck and Bower (1988a) and Estes et al. (1989). We tested the same abstract category structure as was tested in these previous studies; indeed, we used the same sequence of training items and category feedback as did Estes et al. (1989). However, instead of using stimuli composed of additive (present vs. absent) features, as did Estes et al. (1989), we used stimuli composed of substitutive features. One advantage of using substitutive-feature stimuli is that it removes some potential ambiguities that arise during the single-feature test trials. For example, as discussed by Shanks (1990), when additive-feature stimuli are used, the training pattern $f_1f_2^*f_3^*f_4^*$ is simply the feature f_1. During the test trials, when the experimenter presents f_1 alone with no information provided about the values of the other features, it is possible that the subject may confuse the single feature with the *pattern* $f_1f_2^*f_3^*f_4^*$. Estes et al. (1989) attempted to remove this source of confusion during the test trials by filling the missing-feature locations with asterisks. Despite their efforts, we believe there is still room for concern about confusing the single-feature test trials with the single-feature training patterns. If any such confusion existed, the tendency for subjects to estimate $P(A \mid f_1)$ greater than .50 would be easily explained, because the normative $P(A \mid f_1f_2^*f_3^*f_4^*)$ is substantially greater than .50. We argue that by using substitutive features, there is no realistic possibility of confusing the single feature f_1 with the pattern $f_1f_2^*f_3^*f_4^*$, because f_2^*, f_3^*, and f_4^* are positively existing feature values (cf. Gluck & Bower, 1988a, Experiment 3).

In addition to using substitutive-feature stimuli, we extend the Gluck and Bower (1988a) and Estes et al. (1989) studies by collecting a richer set of test-trials data during the transfer phase. In these earlier studies, the researchers collected probability judgments or classification choices of the alternative categories given single features of the training patterns. In Experiment 1, we collect probability judgments and classification choices of the alternative categories given all possible single features, pairs of features, triples of features, and quadruples of features (i.e., the complete patterns). In addition, we test subjects with the null pattern. By collecting this richer data set, we are able to conduct more detailed and rigorous quantitative tests of the competing models than were possible in the Gluck and Bower (1988a) and Estes et al. (1989) studies.

Method

Subjects. The subjects were 144 undergraduates from Indiana University who participated as part of an introductory psychology course requirement. The categorization group consisted of 84 subjects who made classification choices during the test trials, whereas the estimation group consisted of 60 subjects who made direct probability estimates during the test trials.

Stimuli and apparatus. The stimuli were visually displayed charts of four binary-valued symptoms, as follows: stuffy versus runny nose, high versus low blood pressure, diarrhea versus constipation, and muscle relaxation versus muscle tension. The symptoms appeared in a vertically arranged list. Each chart was to be classified into one of two fictional disease categories, *burlosis* versus *midrosis*. The stimulus charts and feedback were presented on the screen of an IBM PC, and the subjects entered their responses on the computer keyboard.

[2] Although attention-weight learning is a fundamental aspect of ALCOVE, it is not critical for explaining performance in this experimental paradigm.

Procedure. The abstract design of the learning phase of Experiment 1 was the same as the one used in the study of Estes et al. (1989, Experiment 1). Subjects learned to classify a sequence of 240 patterns into two probabilistically defined categories. The abstract structure of the sequence was the same for all subjects. The sequence had the property that, over the 240 trials, the base-rate probability of Category A was .25 and of Category B was .75, and the conditional probabilities of the individual features given each category were as described previously. The assignment of feature names to the abstract codings of the stimuli was randomized for each subject, as was the assignment of disease names to Categories A and B. Also, the assignment of dimensions 1 through 4 to lines 1 through 4 of the list was randomly determined for each subject. Corrective feedback was provided on every trial of the learning phase.

As in the Estes et al. study, after every 60 trials of learning, a set of test trials was inserted. During the test trials, single features of the training patterns were presented. Because there were four binary-valued dimensions, there were eight such single-feature test trials. The order of presentation of the single features was randomized for each subject. Subjects in the categorization group classified each feature into either Category A or B, whereas subjects in the estimation group made direct probability estimates of the categories given each feature. Half of the subjects in the estimation group made these probability judgments with respect to Category A (i.e., "What is the probability that a patient with this symptom belongs in Category A?"), and the other half made their judgments with respect to Category B. No feedback was presented during the test trials.

Following the training sequence, a transfer phase was conducted. Subjects were presented, in random order, with all possible single features, pairs of features, triples of features, and quadruples of features, plus the null pattern (i.e., the stimulus with no features). Patterns with mutually exclusive features (e.g., runny nose and stuffy nose) were not presented. Subjects in the categorization group classified each configuration of features into either Category A or B, whereas subjects in the estimation group made direct probability estimates, half with respect to Category A and half with respect to Category B. No feedback was presented during the transfer phase.

Results

Classification learning. The probabilities of correct responses during classification learning, averaged over ten-trial blocks, are displayed separately for the estimation and categorization groups in Figure 3. The choppy appearance of the learning curves arises because the same sequence was used for all subjects, thus the blocks contain items of varying difficulty. The close match between the learning curves of the two groups attests to the reliability of the data.

The competing models were fitted to the learning data on an individual, trial-by-trial basis, by minimizing the sum of squared deviations (SSD) between predicted and observed response probabilities. The best-fitting parameters and summary fits for the models are reported in Table 1. As can be seen in Table 1 the models provide essentially equal fits to the learning data in both groups. Thus, the advantage for the simple adaptive network (the component-cue model) over the context model that was reported by Estes et al. disappears when the context model is elaborated with a memory-decay parameter and a background-noise constant. Additional analyses revealed that the background noise constant played a critical role in allowing the context model to fit these data, whereas the memory-decay parameter played a relatively minor role (Nosofsky, Kruschke, & McKinley, 1991).

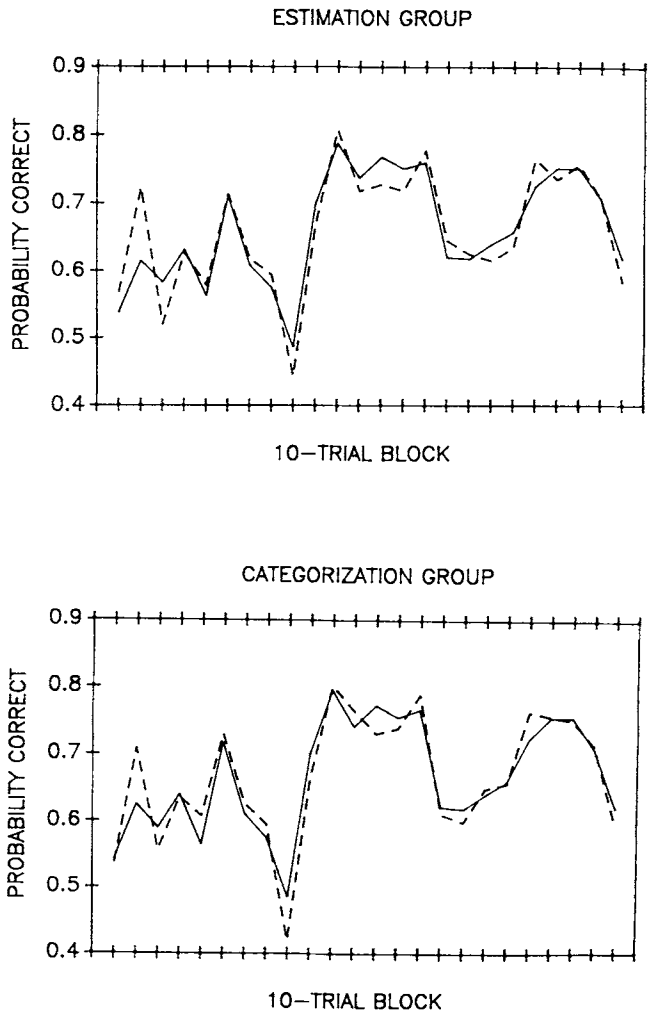

Figure 3. Probabilities of correct classification responses during learning, averaged over ten-trial blocks, in the estimation and categorization conditions. (The dashed lines denote the observed data, and the solid lines give the predictions of the exemplar-based network model [ALCOVE].)

The exemplar-based network model (ALCOVE) fitted the sequence of classification learning data as well as did both the component-cue network and the sequence-sensitive context model (see Table 1). Thus, there is promising support for the idea of integrating an exemplar-based category representation with the learning mechanisms provided by adaptive networks. Figure 3 illustrates the predicted probabilities of correct classification responses for the exemplar-based network model. Although the model was fitted to the data on a trial-by-trial basis, the predictions shown here are averaged over the ten-trial blocks. As can be seen, the fits to the averaged learning data are quite good. The component-cue model and the sequence-sensitive context model achieved roughly the same fits.

In summary, Estes et al. (1989) reported a uniform superiority of Gluck and Bower's (1988a) adaptive network model over an exemplar-based model for predicting the trial-by-trial sequence of classification learning. The finding here is that

Table 1
Fits of the Models to the Trial-by-Trial Learning Data in Experiment 1

Model	SSD	RMSD	%Var	Parameters
Estimation condition				
Component-cue network	2.254	.0969	66.9	$\beta = .022, \beta_b = 0, c = 2.277$
Context	2.135	.0943	68.6	$s = .215, B = .666, T = .0203$
Exemplar-based network (ALCOVE)	2.142	.0945	68.5	$\kappa = 5.223, \beta = .069, c = 1.158$
Categorization condition				
Component-cue network	1.969	.0906	72.4	$\beta = .023, \beta_b = 0, c = 2.344$
Context	1.959	.0903	72.5	$s = .201, B = .632, T = .0192$
Exemplar-based network (ALCOVE)	1.926	.0896	73.0	$\kappa = 5.049, \beta = .072, c = 1.176$

Note. SSD = sum of squared deviations; RMSD = root mean squared deviation; %Var = percentage of variance accounted for. ALCOVE = attentional learning covering map.

when the exemplar model is augmented to include assumptions about learning processes, it fares as well as the Gluck and Bower (1988a) network model at predicting the data. Note that an error-driven learning mechanism is not needed to characterize these learning data because the standard context model with background noise fares as well as the component-cue and exemplar-based networks.

Classification transfer. The complete sets of classification transfer data are reported in Table 2. We begin our discussion of the transfer data by displaying in Figure 4 the subset of results corresponding to single-feature tests and the null pattern. (The averaged data obtained on the single-feature test trials that occurred during the learning phase were essentially the same as the present subset of transfer data, therefore we simply report the transfer data.) The choice function (solid squares) gives the probability with which each item was classified in Category A (the rare category), whereas the estimate function (solid circles) gives the (normalized) probability estimate of Category A membership for each item.[3] To facilitate comparisons, we also display the previous results reported by Gluck and Bower (1988a, Experiment 3) of probability estimates for each item (open circles) and the normative probabilities of Category A given each item (Xs).

The major qualitative result emphasized in the Gluck and Bower (1988a) study was subjects' tendency to estimate that Category A was more probable than Category B given f_1. This key result, which is predicted by the adaptive network model but not by the context model, was replicated in our study. The estimate of Category A given f_1 was not quite as large in our study as reported by Gluck and Bower (1988a, Experiment 3) and was only marginally greater than .50, $t(59) = 1.48$, $p < .10$, one-tailed test. More convincing was that the Category A choice probability for f_1 was extremely large and significantly greater than .50 according to a binomial test ($z = 4.58$, $p < .01$).

Generally, as observed previously by Gluck and Bower (1988a) and Estes et al. (1989), the probability estimates and choices of Category A lie above the normative probabilities. Note that the results of the single-feature tests provide hints that a simple response-bias explanation of this tendency toward base-rate neglect is not tenable. One source of evidence against such an explanation is that the choice probability for f_1^* lies clearly *below* the normative probability. Furthermore, note that the normative probability of Category A given f_1^* is greater than the normative probability of Category A given f_4, whereas the probability estimates reported by Gluck and Bower (1988a, Experiment 3), and the estimates and choice probabilities observed in our study, go in the opposite direction of the normative ones for these two features. Simply "lifting up" the normative curve through multiplication by a bias factor fails to account for this reversal.

Finally, the estimates and choice probabilities of Category A given the null pattern are both below .50. This result provides evidence that subjects have knowledge of the differing base rates for the two categories. Furthermore, this result poses problems for the version of the component-cue model tested by Gluck and Bower (1988a) and Estes et al. (1989). If one faithfully applies the model to predict choice probabilities, then when the network is presented with the null pattern, all dimensional input nodes receive activations of zero, thus the category outputs are both zero. Applying the network-model choice rule (Equation 2) then leads to the prediction that the null pattern is classified in Category A with a probability of .50. By augmenting the component-cue model with the bias node, it is possible to predict the base-rate sensitivity displayed by subjects when presented with the null pattern. Analyses reported by Nosofsky et al. (1991) indicate that the fit of the network model to the transfer data is substantially worse without the bias node.

Theoretical analyses. Our comparisons of the competing models are restricted to the transfer data obtained from the categorization group because there is no generally agreed-on method for using the models to predict direct probability estimates. The models were fitted to the transfer data by searching for the parameters that minimized the SSD between predicted and observed response probabilities over all 81 patterns.[4] Following Estes et al. (1989) in fitting the context

[3] The normalized estimate was computed by dividing the average estimate for Category A by the sum of the average estimates for Categories A and B.

[4] In these analyses, we allowed the parameters to vary freely rather than constraining them to be the same as the best-fitting learning parameters. In other analyses (reported by Nosofsky, Kruschke, & McKinley, 1991), we fitted the models simultaneously to the learning and transfer data with all parameters held fixed. The analyses involving simultaneous fits led to the same conclusions as those reported here involving separate fits.

Table 2
Normative, Estimated, and Observed and Predicted Choice Probabilities of Category A (the Rare Category) for Each of the 81 Transfer Patterns

Pattern	Normative	Estimated	Observed choice	Predicted choice
Null				
0000	.250	.364	.119	.226
Single				
1000	.500	.558	.750	.586
2000	.143	.269	.036	.075
0100	.308	.366	.465	.364
0200	.222	.361	.262	.216
0010	.200	.403	.346	.214
0020	.280	.372	.334	.274
0001	.100	.324	.215	.120
0002	.400	.489	.524	.441
Double				
1100	.571	.520	.679	.628
1200	.462	.477	.512	.500
1010	.429	.498	.500	.543
1020	.538	.488	.595	.616
1001	.250	.478	.429	.388
1002	.667	.526	.750	.748
2100	.182	.282	.155	.182
2200	.125	.201	.107	.094
2010	.111	.252	.095	.079
2020	.163	.285	.095	.101
2001	.053	.189	.024	.045
2002	.250	.314	.155	.181
0110	.250	.387	.381	.348
0120	.341	.358	.393	.402
0101	.129	.308	.310	.241
0102	.471	.424	.524	.534
0210	.176	.326	.179	.210
0220	.250	.364	.203	.257
0201	.087	.299	.119	.136
0202	.364	.450	.393	.379
0011	.077	.331	.143	.120
0012	.333	.417	.441	.408
0021	.115	.328	.262	.155
0022	.438	.410	.524	.482
Triple				
1110	.500	.470	.560	.594
1120	.609	.511	.607	.643
1101	.308	.408	.453	.478
1102	.727	.517	.750	.739
1210	.391	.400	.465	.470
1220	.500	.483	.548	.530
1201	.222	.428	.358	.355
1202	.632	.531	.703	.645
1011	.200	.409	.286	.364
1012	.600	.531	.655	.705
1021	.280	.401	.453	.429
1022	.700	.566	.726	.758
2110	.143	.304	.238	.185
2120	.206	.298	.226	.216
2101	.069	.252	.083	.127
2102	.308	.355	.310	.310
2210	.097	.264	.071	.099
2220	.143	.286	.072	.121
2201	.045	.215	.071	.066
2202	.222	.306	.119	.188
2011	.040	.255	.083	.050
2012	.200	.301	.179	.179
2021	.061	.229	.119	.064
2022	.280	.327	.238	.220
0111	.100	.333	.203	.238
0112	.400	.395	.488	.506
0121	.147	.320	.226	.278
0122	.509	.466	.500	.558
0211	.067	.290	.143	.139
0212	.300	.368	.369	.359
0221	.100	.334	.155	.170
0222	.400	.404	.393	.416
Complete				
1111	.250	.405	.429	.456
1112	.667	.547	.655	.704
1121	.341	.419	.393	.503
1122	.757	.592	.691	.742
1211	.176	.421	.274	.341
1212	.563	.586	.560	.608
1221	.250	.485	.357	.392
1222	.667	.613	.667	.659
2111	.053	.287	.131	.134
2112	.250	.328	.262	.304
2121	.080	.243	.083	.156
2122	.341	.368	.262	.343
2211	.034	.220	.059	.073
2212	.176	.310	.202	.189
2221	.053	.214	.071	.087
2222	.250	.329	.155	.222

Note. For patterns, 1 = positive feature, 2 = negative feature, and 0 = missing feature. The predicted choice probabilities are for the elaborated exemplar-based network (ALCOVE: attentional learning covering map).

model and ALCOVE, we computed similarity only over those dimensions that had nonmissing values.

The least-squares parameters and summary fits for the models are reported in Table 3. As can be seen, the sequence-sensitive context model lags behind the (biased) component-cue network and the exemplar-based network in its quantitative fits to the transfer data. The latter models perform about equally well, with a slight advantage to the (biased) component-cue network.

To test for the potential role of differential dimension salience, we also fitted elaborated versions of each model to the transfer data. In the elaborated component-cue model, separate learning rates were allowed for each input dimension. In the context model and ALCOVE, separate similarity or attention-weight parameters were allowed for each dimension. The fits of these models are shown with those of the baseline models in Table 3. It is not surprising that adding these free parameters improved the fits of all the models. However, the context model clearly still lags behind the component-cue network and ALCOVE at predicting the transfer data.

A plausible hypothesis is that the shortcomings of the sequence-sensitive context model result from its use of a constant-increment learning mechanism rather than the error-driven learning mechanism found in the network models. This hypothesis is supported by examination of the specific predictions of the baseline version of the sequence-sensitive context model. First, note that the context model's best-fitting similarity parameter (s) was zero and the best-fitting decay rate (T) was zero. With these parameters, the predictions of the context model are roughly the normative probabilities of

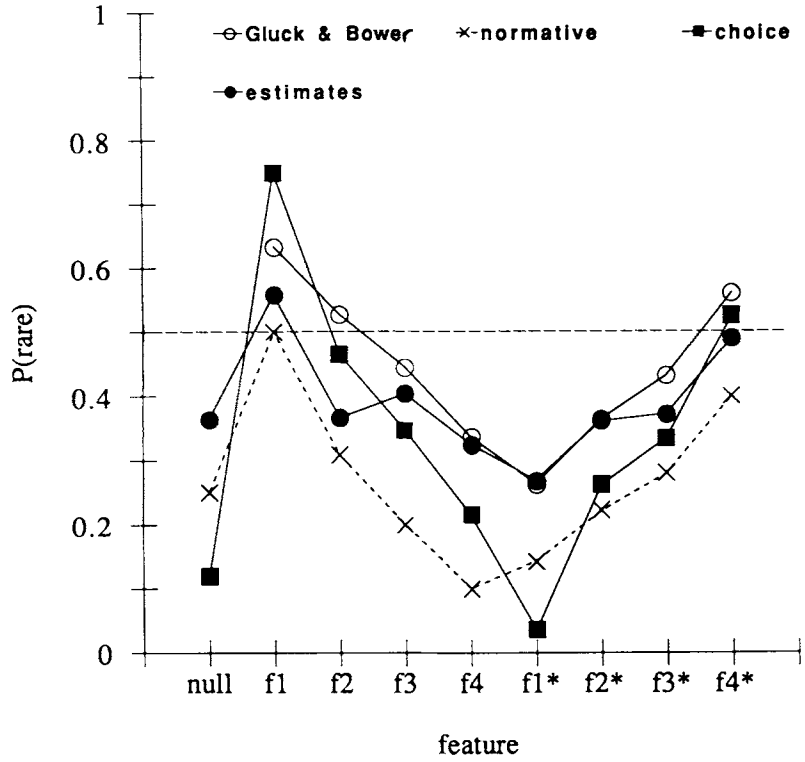

Figure 4. Transfer results for the single-feature and null-pattern tests. (Solid squares denote Category A choice probabilities, and solid circles denote Category A probability estimates. Open circles denote the Category A probability estimates reported previously by Gluck and Bower [1988a, Experiment 3], whereas Xs denote the normative probabilities of Category A.)

the categories, given each combination of features. (The background constant serves to slightly pull all of the normative predictions toward .50.) Thus, the model tended to underpredict the Category A choice probabilities for the single-feature tests illustrated in Figure 4 and to severely underpredict the Category A choice probability for feature f_1. An extended version of the model with a response-bias parameter fitted the data only slightly better than the baseline version, and worse than the three-parameter network models (SSD = .610). Thus, the base-rate neglect phenomenon exhibited in Experiment 1 cannot be explained by the context model solely in terms of an overall response bias toward Category A. The results of this model-based analysis agree with our earlier observations regarding the reversal for features f_1^* and f_4, wherein the normative probability of Category A is greater given than f_4, but the observed data go in the opposite direction.

Predicting base-rate neglect. The predictions of the exemplar-based network (ALCOVE) are shown with the ob-

Table 3
Fits of the Models to the Transfer Data in Experiment 1 (Categorization Condition Only)

Model	SSD	RMSD	%Var	Parameters
Baseline				
Component-cue network	.488	.0969	86.0	$\beta = .009$, $\beta_b = .024$, $c = 1.954$
Context	.669	.0909	80.8	$s = 0$, $B = 3.12$, $T = 0$
Exemplar-based network (ALCOVE)	.530	.0809	84.0	$\kappa = 1.879$, $\beta = .024$, $c = 1.713$
Elaborated				
Component-cue network	.213	.0513	93.9	$\beta_1 = .010$, $\beta_2 = .012$, $\beta_3 = .003$ $\beta_4 = .005$, $\beta_b = .045$, $c = 4.875$
Context	.450	.0746	87.1	$s_1 = 0$, $s_2 = 0$, $s_3 = .611$, $s_4 = .136$, $B = 1.318$, $T = 0$
Exemplar-based network (ALCOVE)	.210	.0510	94.0	$\alpha_1 = .279$, $\alpha_2 = .691$, $\alpha_3 = .165$, $\alpha_4 = .205$, $\beta = .022$, $c = 3.769$

Note. SSD = sum of squared deviations; RMSD = root-mean-squared deviation; %Var = percentage of variance. ALCOVE = attentional learning covering map.

served data in Table 2. In contrast with the context model, ALCOVE predicts the clear tendency for subjects to classify f_1 in Category A with probability greater than .50. The model is also able to predict the reversal involving f_4 and f_1^*. In general, ALCOVE predicts the tendency for subjects' choice probabilities on the single-feature tests to exceed the normative ones, although it fails to predict the magnitude of the effect. Finally, the model predicts the base-rate sensitivity observed for the null pattern.

ALCOVE predicts apparent base-rate neglect for the single-feature tests by virtue of its error-driven learning and similarity-activated exemplar representation. Consider, for the sake of exposition, only two features. Then there is a two-dimensional stimulus space, with four stimuli: (f_1, f_2), (f_1, f_2^*), (f_1^*, f_2), and (f_1^*, f_2^*). Figure 5 shows the frequency of each stimulus as a member of the rare or common category. (The feature frequencies illustrated in Figure 5 are chosen for the sake of exposition in this two-dimensional example.) Note that the marginal probability of the rare category given feature f_1 is 15/30, or .50, thus a Bayesian classifier would not favor either the rare or the common category, given f_1 alone.

Figure 5 acts as a geometric analogue of the two-dimensional stimulus space. In the exemplar-based version of ALCOVE, one can imagine four hidden nodes centered on the four cells of Figure 5. Consider the two neighboring cells in the lower row, $f_1^*f_2^*$ and $f_1f_2^*$. Both are more likely, given the *common* category (50 of 55 cases, and 14 of 19 cases, respectively). Feature-pair $f_1^*f_2^*$ is the more frequent pair of the two (55 cases vs. 19), therefore the hidden node centered on it will develop a strong positive connection to the common category node (and a negative connection to the rare category node). When the neighboring feature-pair $f_1f_2^*$ occurs, the hidden node over $f_1^*f_2^*$ is partially activated, in turn activating the common category node. As a result, relatively little error is produced, and the connection from the $f_1f_2^*$ node becomes only *weakly* weighted to the common category node. On the other hand, the node centered on the top left cell, f_1f_2, must develop a moderately strong positive weight to the rare category node, forced to be even stronger by the presence of conflicting neighbors $f_1f_2^*$ and $f_1^*f_2$.

f1,f2 10 / 1 (11)	f1*,f2 5 / 14 (19)	f2 15 / 15 (30)
f1,f2* 5 / 14 (19)	f1*,f2* 5 / 50 (55)	f2* 10 / 64 (74)
f1 15 / 15 (30)	f1* 10 / 64 (74)	

Figure 5. Each cell shows the frequency with which the corresponding stimulus is in each category: rare frequency/common frequency (total frequency). (The marginal frequencies of individual features are also shown.)

When the single feature f_1 is input to ALCOVE, the hidden nodes centered over the left column, f_1f_2 and $f_1f_2^*$, are maximally activated, and the nodes centered over the right column are only partially activated. Because the connection weights from f_1f_2 strongly favor the rare category, and the connection weights from $f_1f_2^*$ only weakly favor the common category, the result is that the rare category node is more strongly activated than the common category node, and ALCOVE appears to exhibit base-rate neglect.

In summary, it is the dual effect of error-driven learning and similarity-based hidden node activations that lets ALCOVE exhibit base-rate neglect. If the learning rule were not error driven, as in the context model, or if similar hidden nodes were not co-activated, as could happen in ALCOVE with extremely high sensitivity values (κ in Equation 15), then exemplars would have no influence on each other's learning, and base rates would not be neglected.[5]

Summary and Discussion

In Experiment 1 we repeated the probabilistic classification learning paradigm used previously by Gluck and Bower (1988a, Experiment 3), except, following Estes et al. (1989), we used a fixed sequence of training items for all subjects and collected choice data in addition to estimation data during the transfer tests. We also extended these earlier studies by testing subjects with the complete powerset of features during transfer.

Various of the key phenomena that were observed in these previous studies were replicated. Most notably, we obtained evidence of a tendency toward base-rate neglect when subjects made classification judgments for single features of multiple-feature patterns. Because we used substitutive-feature stimuli, we can rule out the idea that the phenomenon was the result of subjects' confusing the single-feature stimuli with complete patterns that had one feature present and three features absent (cf. Gluck & Bower, 1988a, Experiment 3; Shanks, 1990).

By augmenting the standard context model with assumptions about learning processes, we were able to achieve as good a fit to the sequence of learning data as was achieved by the adaptive network models. However, even this augmented context model fared worse than the network models at quantitatively predicting the transfer data. Moreover, adding a response-bias parameter to the context model in an attempt to account for base-rate neglect did not allow it to characterize the complete set of transfer data as well as did the network models.

Nevertheless, an alternative exemplar-based model, namely the exemplar-based version of ALCOVE, fitted both the learning and transfer data as well as did the component-cue net-

[5] ALCOVE predicts base-rate sensitivity for the null pattern as follows: When the null pattern is presented, all exemplar nodes in the network are maximally activated. Because of the higher base rate of Category B, the association weights pointing to the B output node tend to be larger than those pointing to the A output node, thus, the tendency to choose Category B.

work model. The exemplar-based version of ALCOVE shares with the component-cue model the assumption of an error-driven, interactive learning rule, in which basic units stored in memory "compete" with one another to become associated with the alternative categories. Unlike the component-cue model, however, in which the memory representation consists of connections between individual feature values and the categories, the memory representation in ALCOVE is exemplar-based. Patterns activate individual exemplar nodes stored in memory according to the same similarity rule assumed in the context model, and the model learns associations between these exemplar nodes and the categories.

Unfortunately, the category structure that was tested in Experiment 1 (the same one used by Gluck and Bower, 1988a, and Estes et al., 1989) was not highly diagnostic for discriminating between the component-cue network and the exemplar-based network. The reason is that the categories were defined over independent probability distributions of the values on the component dimensions. A variety of competing models, including the context model and certain independent feature-frequency models, make formally identical predictions for such structures (Estes, 1986a; Nosofsky, 1990). More diagnostic category structures can be designed by introducing interdimensional correlations. In Experiment 2, we contrast the competing models on their ability to predict performance in a category learning paradigm involving a structure with interdimensional correlations.

Experiment 2

Whereas the focus in Experiment 1 was on the nature of the learning rule, the focus in Experiment 2 is on the nature of the category representation. Nosofsky (in press-a) demonstrated that the component-cue network model is in essence a multiplicative-similarity prototype model (for related theoretical analyses, see Golden & Rumelhart, 1989, and Massaro & Friedman, 1990). We define *prototype* here as some single point in the multidimensional space in which the category exemplars are embedded. Let $p_1 = (p_{11}, p_{12}, \ldots, p_{1m})$ be the Category 1 prototype, where p_{1m} denotes the psychological value of prototype 1 on dimension m, and likewise for the Category 2 prototype. Let $S(x, p_1)$ denote the overall similarity between input pattern x and prototype 1. In the multiplicative-similarity prototype model, the similarity of an input pattern to the prototype is computed by using the same multiplicative rule as is used in the context model:

$$S(x, p_1) = \prod s(x_m, p_{1m}), \tag{17}$$

where $s(x_m, p_{1m})$ is the similarity of x to prototype 1 on dimension m, and likewise for $S(x, p_2)$. The component-cue model is a special case of the following multiplicative-similarity prototype model:

$$\begin{aligned} P(R_1 \mid x) &= S(x, p_1)/[S(x, p_1) + S(x, p_2)] \\ &= \prod s(x_m, p_{1m})/[\prod s(x_m, p_{1m}) \\ &\quad + \prod s(x_m, p_{2m})]. \end{aligned} \tag{18}$$

(See Nosofsky, in press-a, for a proof.) Although not made explicit in the notation, the values of the dimensional similarity parameters in Equation 18 will vary from trial to trial. The values of these parameters are determined by the weights in the network, which are learned trial by trial by the delta rule.

In numerous studies, Medin and his associates (e.g., Medin, Altom, & Murphy, 1984; Medin, Dewey, & Murphy, 1983; Medin & Schaffer, 1978; Medin & Smith, 1981) and Nosofsky (1987, 1988, 1991) have systematically compared the quantitative predictions of prototype models with those of exemplar models (the context model). The outcomes of these comparisons have been overwhelmingly in favor of the context model (see Nosofsky, in press-a, for a review). However, the focus of these previous studies was on the ability of these models to quantitatively predict performance in classification transfer situations, following the completion of an initial learning phase. In Experiment 2, we use one of the diagnostic category structures tested previously by Medin but compare the competing models on their ability to predict the details of classification *learning*. We also test the models' ability to predict transfer performance at various stages of the learning sequence.

The category structure that is tested is shown in Table 4. The stimuli vary along four binary-valued dimensions. Members of Category A tend to have (logical) value 1 on each of their dimensions, whereas members of Category B tend to have (logical) value 2. Note that the categories are linearly separable, which is a necessary and sufficient condition for accurate classification with a (weighted dimensions) prototype strategy. Thus, if subjects' natural strategy is to learn category prototypes, the strategy would succeed for this category structure.

As in Experiment 1, all subjects were presented with the same fixed sequence of training items, consisting of Stimuli 1 through 9 in Table 4. Following each of four training blocks, a transfer phase was conducted in which all 16 stimuli shown in Table 4 were presented. Corrective feedback was provided during the training blocks but was withheld during each

Table 4
Category Structure Tested in Experiment 2

Structure	Pattern	Dimension 1	2	3	4
Category A	1	1	1	1	2
	2	1	2	1	2
	3	1	2	1	1
	4	1	1	2	1
	5	2	1	1	1
Category B	6	1	1	2	2
	7	2	1	1	2
	8	2	2	2	1
	9[b]	2	2	2	2
Transfer test patterns	10	1	2	2	1
	11	1	2	2	2
	12[a]	1	1	1	1
	13	2	2	1	2
	14	2	1	2	1
	15	2	2	1	1
	16	2	1	2	2

Note. The category structure used in this experiment is from Medin and Schaffer (1978). [a] Category A prototype. [b] Category B prototype.

transfer phase. The theoretical goal was to predict quantitatively the trial-by-trial sequence of classification learning and the evolution of transfer performance.

Because the component-cue model is a multiplicative-similarity prototype model, it makes the strong prediction that the prototype of Category A [1111] (Pattern 12 in Table 4), which is never presented during training, will be classified in Category A with probability at least as high as any of the other patterns, including the old exemplars. By contrast, both the context model and the exemplar-based network model can predict that various of the old exemplars will be classified in Category A with higher probability than the prototype. (Because Prototype B [2222] is an actual training exemplar, all models tend to predict that it will be classified in Category B with very high probability.)

A second fundamental contrast between the component-cue model and the exemplar models regards their predictions of performance on Training Patterns 1 and 2 of Category A. Intuitively, Training Pattern 1 is at least as similar to the Category A prototype as is Training Pattern 2. (The patterns match on all dimensions except Dimension 2, and on this dimension Pattern 1 matches the A prototype, whereas Pattern 2 mismatches.) Thus, the component-cue model predicts that during both learning and transfer, Pattern 1 will be classified in Category A with higher probability than Pattern 2. (We verify this prediction formally in the Theoretical Analyses section.) By contrast, both the context model and the exemplar-based network model tend to predict an advantage for Pattern 2 over Pattern 1. The reason is that Pattern 1 is highly similar to only one other exemplar in its own category and is highly similar to two exemplars from the contrast category, whereas Pattern 2 is highly similar to two other exemplars in its own category and is not highly similar to any exemplars in the contrast category (see Medin & Schaffer, 1978, for a more extended discussion).

Method

Subjects. The subjects were 40 undergraduates from Indiana University who participated as part of an introductory psychology course requirement.

Stimuli and apparatus. The stimuli were geometric forms with lines that filled their interiors. The stimuli varied along four binary-valued dimensions: size (large or small), shape (triangles or squares), type of interior lines (dotted or dashed), and density of interior lines (high or low). Dimensions 1 through 4 in Table 4 corresponded to line type, shape, size, and line density, respectively. The stimuli were presented on the screen of an IBM PC, and subjects entered responses by using the computer keyboard.

Procedure. The category structure shown in Table 4 was used. The learning sequence was organized into 4 blocks of 63 trials each. During each block, each of the nine training items was presented seven times. Order of presentation of the items was randomized. The same sequence was used for all subjects. On each trial, subjects judged whether the item belonged to Category A or B, and corrective feedback was then provided.

Following each learning block, a transfer phase was conducted. During each of the four transfer blocks, all 16 stimuli shown in Table 4 were presented in a newly randomized order for each subject. Subjects judged whether each item belonged to Category A or B. No feedback was provided.

Results

The key qualitative result of interest in the learning phase is the ordering of difficulty for Patterns 1 and 2 of Category A. Figure 6 plots the probability of correct classifications for these two patterns as a function of block of learning. Performance on both patterns improved as a function of learning, with Pattern 2 being classified more accurately than Pattern 1 throughout, $t(39) = 5.04$, $p < .01$. The superior learning performance for Pattern 2 is consistent with the predictions of the context model and the exemplar-based network model but contradicts the predictions of the component-cue model.

The complete set of transfer data is shown in Table 5. The table shows the probability with which each pattern was classified in Category A during each of the four transfer blocks. A key result of interest concerns performance on the Category A prototype compared with the old exemplars of Category A. The component-cue model predicts that the prototype will be classified in Category A with probability at least as high as *any* of the old exemplars, whereas the context model and the exemplar-based network model can predict advantages for various of the old exemplars. Figure 7 plots, for each of the four transfer blocks, the probability with which the prototype was classified in Category A, and the average probability with which the old A exemplars were correctly classified in their category. As shown in Figure 7, not only was the prototype not the best classified pattern but it was classified in Category A with probability less than the average of the old exemplars. Indeed, the proportion of errors on Patterns 2 and 3 (which are predicted by the exemplar-based models to be the best classified, old A patterns) was significantly less than that of the A prototype, $t(39) = 2.06$, $p < .05$. These results strongly contradict the predictions of the component-cue model but can be explained by the context model and the exemplar-based network model.[6]

Theoretical analyses. Before reporting the results of the theoretical analyses, we introduce an augmented version of the context model that we tested in Experiment 2. Quantitative analyses of identification confusion data conducted by Nosofsky (1987) provided evidence that similarities among perceptual objects may decrease as a function of learning. This result agrees with the classic idea that perceptual differentiation among objects increases with experience (e.g., Gibson & Gibson, 1955). We hypothesized that the context model's predictions of classification learning might improve if the similarity parameters in the model were allowed to decrease as a function of learning (see also Estes, 1986b). To implement this idea, we assumed that the psychological distance (D) between mismatching values on each dimension

[6] We emphasize that, had a prototype enhancement effect been observed, it would not have been inconsistent with the predictions of the exemplar models. As discussed extensively in previous work (e.g., Hintzman, 1986; Medin & Schaffer, 1978), exemplar models can predict advantages for either the prototype or the old exemplars, depending on similarity relations that hold under particular experimental conditions. The poor performance on the prototype in Experiment 2, however, cannot be explained by the component-cue network model.

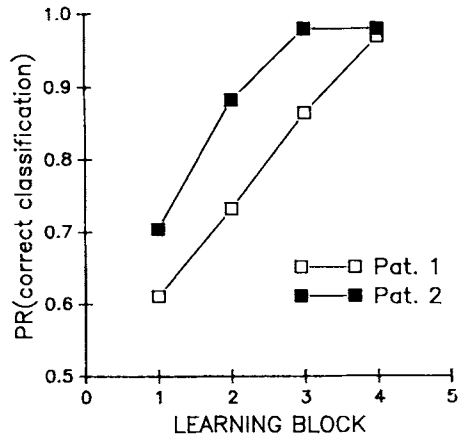

Figure 6. Probability of correct classifications for Pattern 1 (open squares) and Pattern 2 (solid squares) as a function of block of learning.

grew linearly with trials of learning (t), $D = a + b \cdot t$, where a and b are freely estimated parameters. Similarity was assumed to be an exponential decay function of distance, $s = \exp(-D)$. Thus, on each subsequent trial of learning, the similarity parameter that enters into the response rule decreases. We again used the background-noise constant (B) in modeling the data, and assumed that its value decreased exponentially with trials at the same rate that similarity decreased (b). The memory decay-rate parameter was held fixed at $T = 0$. Similar to the competing models, this version of the context model has three free parameters (a, b, and B). Versions of ALCOVE can also be fitted in which similarity is adjusted by error-driven learning (see Kruschke, 1990a), but we did not test these models in this research.

Quantitative predictions of learning and transfer. We start the quantitative comparisons by focusing on the transfer data because these data provide the most diagnostic information for discriminating among the models. The models were fitted to the data by searching for the parameters that minimized SSD computed over all 16 patterns in each of the four transfer blocks. The fits of the context model, the exemplar-based network model (ALCOVE), and the component-cue model to each transfer block are reported in Table 6. As can be seen, the component-cue model fares extremely poorly, with a total SSD approximately four times as great as both the context model and the exemplar-based network model. The fits of the context model and the exemplar-based network model are quite good, particularly in the latter two transfer blocks.

To provide insight into the results of these quantitative comparisons, Figure 8 plots the Category A response probabilities for each of the 16 transfer patterns (averaged over the four transfer blocks), together with the predictions of the models. Both the context model and the exemplar-based network model perform remarkably well. The component-cue model, however, is far off on many of its predictions. As discussed earlier, the model predicts that the prototype of Category A (Pattern 12) will be classified in Category A with higher probability than any of the old exemplars, but this result was not observed. The component-cue model also underpredicts correct classification for many of the old exemplars (Patterns 2, 4, 6, 7, and 8). Finally, it misorders the Category A response probabilities for Patterns 1 and 2 of Category A, as we explained earlier. By contrast, the context model and the exemplar-based network model account accurately for these major phenomena.

Figure 9 provides perspective on the ability of the competing models to predict the evolution of transfer performance over blocks. We again plot the average Category A response probabilities for the old A exemplars and the A prototype for each block of transfer, and also show the predictions of the models. Both the context model and the exemplar-based network account for these data fairly well, but the component-cue model severely mispredicts the data. Not only are its quantitative predictions poor but it misorders the relative difficulty for the old exemplars and the prototype throughout the entire course of learning.

Table 5
Category A Response Probabilities for Each of the Patterns During the Transfer Blocks of Experiment 2

Structure	Pattern	Block 1	2	3	4	Average
Old A exemplars	1	.53	.93	.97	.93	.840
	2	.78	.90	1.00	1.00	.920
	3	.75	.95	1.00	1.00	.925
	4	.82	.85	.95	1.00	.905
	5	.70	.60	.88	.93	.777
Old B exemplars	6	.35	.05	.18	.00	.130
	7	.35	.25	.10	.15	.213
	8	.20	.07	.03	.00	.075
	9	.25	.12	.00	.05	.105
Untrained patterns	10	.62	.65	.62	.65	.635
	11	.53	.42	.42	.42	.447
	12[a]	.70	.82	.90	.90	.830
	13	.45	.45	.55	.45	.475
	14	.75	.40	.50	.60	.563
	15	.53	.45	.65	.62	.563
	16	.23	.17	.20	.12	.180

[a] Category A prototype.

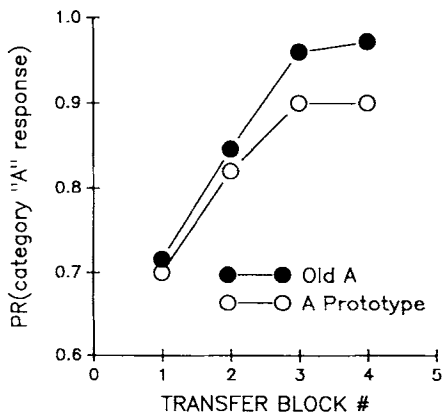

Figure 7. Average probability of correct classifications for the old exemplars of Category A (solid circles) and the Category A prototype (open circles) as a function of block of transfer.

Table 6
Fits of the Models to the Transfer Data in Experiment 2

Model	Block 1	Block 2	Block 3	Block 4	Total	Parameters
Baseline						
Component-cue network	.325	.327	.428	.496	1.575	$\beta = .013$, $\beta_b = .003$, $c = 1.536$
Context	.115	.162	.050	.024	.350	$a = 1.039$, $b = .0105$, $B = 5.104$
Exemplar-based network (ALCOVE)	.172	.159	.037	.026	.394	$\kappa = 6.895$, $\beta = .046$, $c = 2.255$
Configural-cue	.180	.177	.130	.106	.593	$\beta = .0039$, $\beta_b = .016$, $c = 1.627$
Elaborated						
Component-cue network	.314	.325	.424	.491	1.554	$\beta_1 = .014$, $\beta_2 = .006$, $\beta_3 = .012$, $\beta_4 = .016$, $\beta_b = .006$, $c = 1.499$
Context	.116	.159	.051	.025	.350	$a_1 = 1.075$, $a_2 = 1.025$, $a_1 = 1.025$, $a_4 = 1.012$, $b = .0105$, $B = 5.119$
Exemplar-based network	.140	.168	.045	.026	.379	$\alpha_1 = 1.645$, $\alpha_2 = 1.598$, $\alpha_3 = 1.639$, $\alpha_4 = 2.099$, $\beta = .046$, $c = 2.150$
Configural-cue (Version 1)	.147	.136	.035	.024	.342	$\beta_1 = 0$, $\beta_2 = .0035$, $\beta_3 = .0009$, $\beta_4 = .0207$, $\beta_b = .022$, $c = 2.053$
Configural-cue (Version 2)	.152	.140	.075	.047	.415	$\beta = .0009$, $\gamma_1 = 2.063$, $\gamma_2 = 2.612$, $\gamma_3 = 1.919$, $\gamma_4 = 1.314$, $\beta_b = .0022$, $c = 1.573$

Note. Fits of models are expressed as sums of squared deviations (SSDs). ALCOVE = attentional learning covering map.

The fits of the competing models to the trial-by-trial sequence of learning data are reported in Table 7. Again, the SSD for the component-cue model far exceeds that for the context model and the exemplar-based network. The main failing of the component-cue model with respect to its predictions of learning was its misordering of difficulty for Patterns 1 and 2. We show in Table 8 the fits of the competing models when the parameters are constrained to be constant across learning and transfer. The failings of the component-cue model relative to the context model and the exemplar-based network are even more dramatic than before.

Because the assignment of physical dimensions to the logical category structure was held fixed in Experiment 2, it is important to test elaborated versions of the models that allow for differential salience of the dimensions. Thus, we fitted a version of the component-cue model in which the learning rate on each dimension was allowed to be a free parameter. As shown in Tables 6, 7, and 8, for both the learning and transfer data, even this elaborated model with six free parameters performed far worse than the three-parameter context model and the three-parameter exemplar-based network. We also tested versions of the context model and the exemplar-based network model that had additional free parameters. For both models, differential similarity parameters (or attention weights) were allowed for each dimension. It is surprising that adding these free parameters led to virtually no improvement in the fit of either model. We consider this result to be fortuitous and expect that, in most experimental situations, differential similarities across dimensions will be needed to adequately model the data, as has been found in previous work (e.g., Kruschke, 1992; Medin & Smith, 1981; Nosofsky, 1984, 1987, 1989).

Tests of Gluck and Bower's (1988b) configural-cue model. Although the main purpose of this research was to compare and contrast the context model, the component-cue model, and the exemplar-based network model, we also conducted preliminary tests of Gluck and Bower's (1988b) configural-cue model. As discussed previously, this model is similar to the component-cue model, except that instead of having the input nodes code individual feature values, the input nodes code all possible configurations of features. For example, if the network were presented with a large black square, individual feature nodes coding large, black, and square would be activated, but so would nodes coding the pairwise configurations *large-black*, *large-square*, and *black-square*, as well as a node coding the three-way configuration *large-black square*. In Experiment 2 the stimuli varied along four binary-valued dimensions, which means there are 80 input nodes, 15 of which are activated on each stimulus presentation.[7] (In addition, a bias node with a separate learning rate was included.) In all other respects, the configural-cue network operates in the same manner as the component-cue network.

The configural-cue network is similar to the exemplar-based network in that both have exemplar nodes that are activated when patterns are presented. (In the configural-cue model, the exemplar nodes are the nodes that code the complete configurations.) Unlike the exemplar-based network, however, the configural-cue model includes all the lower order nodes (i.e., the single, double, and triple nodes). The configural-cue network also differs from the exemplar-based network in that activation of each of the input nodes is all-or-none, whereas in the exemplar-based network, activation of each exemplar node is proportional to its similarity to the input pattern. Finally, the exemplar-based network has mech-

[7] There are 8 single nodes (4 single dimensions × 2 values on each dimension), 24 double nodes (4-choose-2 pairwise dimensional combinations × 2^2 value combinations), 32 triple nodes (4-choose-3 three-way dimensional combinations × 2^3 value combinations), and 16 quadruple nodes (4-choose-4 four-way dimensional combinations × 2^4 value combinations).

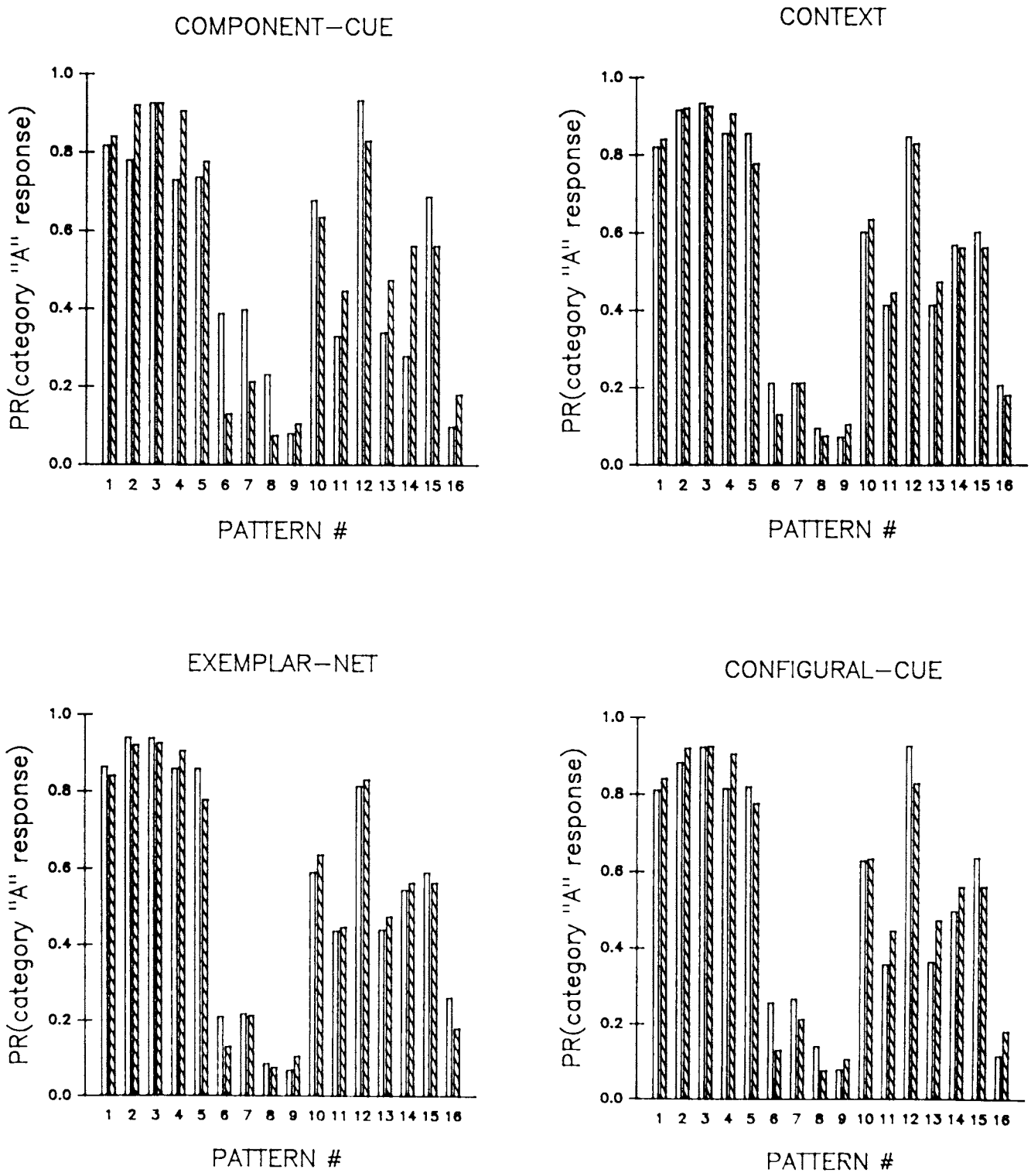

Figure 8. Average Category A response probabilities for the sixteen transfer patterns (cross-hatched bars), plotted with the predictions of the alternative models (open bars).

anisms for learning attention weights for individual dimensions, whereas such mechanisms are not present in the configural-cue model (cf. Kruschke, 1990b, pp. 16–17, 24–26, 44–47).

The configural-cue model was fitted to our data in the same manner as described for the other models. The baseline version of the model had the same three free parameters as the component-cue model: the learning rate β, the scale parameter

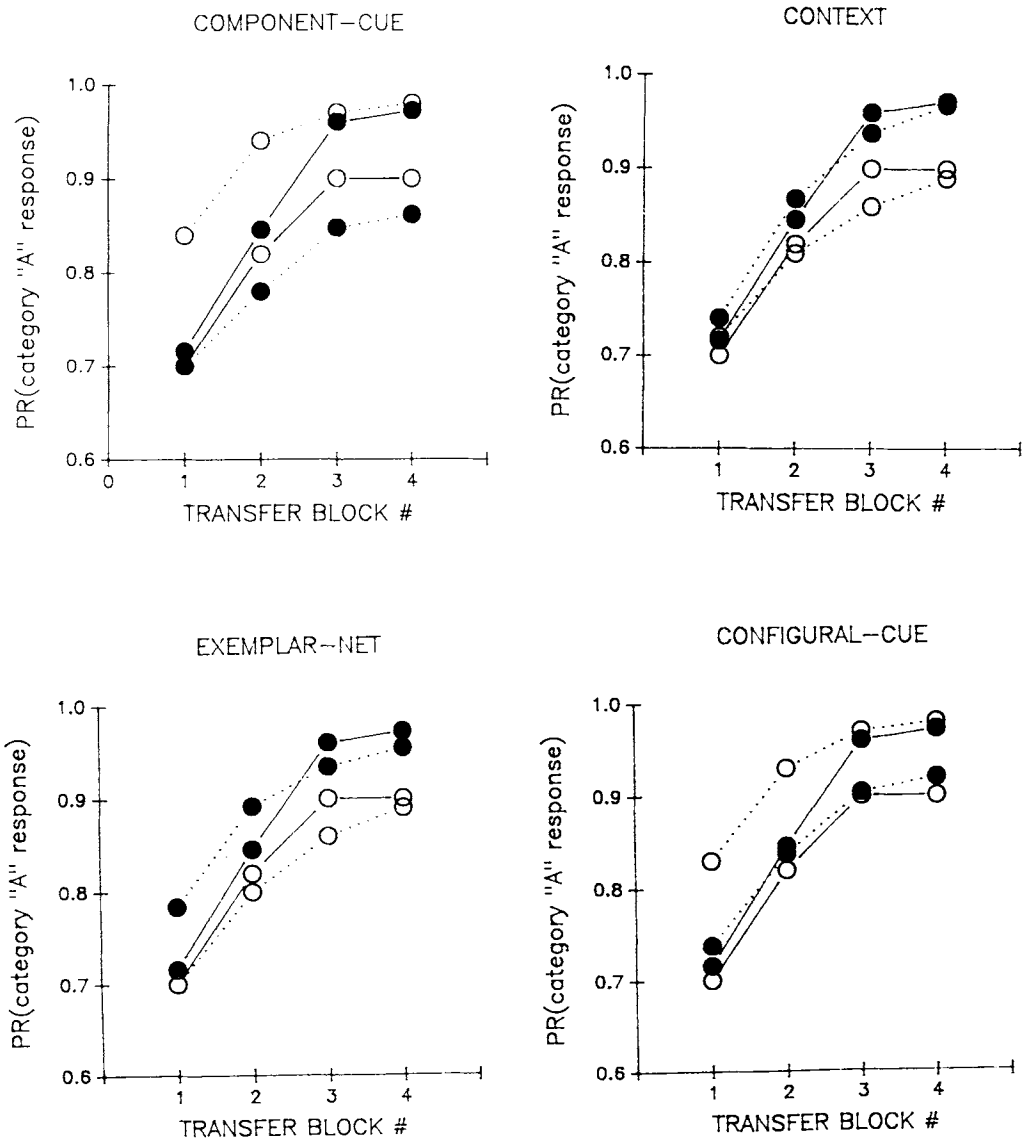

Figure 9. Observed and predicted Category A response probabilities for the old A exemplars (solid circles and the A prototype (open circles), plotted as a function of block of transfer. (Solid lines denote observed probabilities, and dotted lines denote predicted probabilities.)

c, and a learning rate β_b for a bias node. The results of the analyses are presented along with those of the other models in Tables 6, 7, and 8, and graphical presentations of the model's predictions are shown along with those of the other models in Figures 8 and 9. The baseline version of the configural-cue model performed far better than the component-cue model at predicting both the learning and transfer data, and performed as well as the context model and exemplar-based network at predicting the learning data. However, the baseline configural-cue model clearly performed worse than the context model and the exemplar-based network model at predicting the transfer data. The main problem with the model was that, similar to the component-cue model, it predicted performance on the Category A prototype that was much too high. Indeed, in a similar manner to the component-cue model, it predicted that the prototype would be classified into Category A with higher probability than *any* of the old exemplars (see Figures 8 and 9). The configural-cue model makes this prediction because the single-feature nodes apparently overtake the exemplar nodes in influencing the total category outputs.

We also fitted a variety of alternative versions of the configural-cue model to the learning and transfer data (see the Appendix). Basically, by adding free parameters to the configural-cue model, we were able to find versions that performed as well as, but no better than, the context model and ALCOVE. These free parameters had the effect of placing almost all of the learning on the higher order configuration nodes (i.e., the triple nodes and the exemplar nodes). This type of learning process is the one assumed in the context

Table 7
Fits of the Models to the Learning Data in Experiment 2

Model	SSD	Parameters
Baseline		
Component-cue network	4.417	$\beta = .035, \beta_b = 0, c = 2.143$
Context	2.961	$a = 1.203, b = .0081, B = .780$
Exemplar-based network (ALCOVE)	3.030	$\kappa = 5.193, \beta = .160, c = 1.406$
Configural-cue	3.065	$\beta = .0098, \beta_b = 0, c = 1.443$
Elaborated		
Component-cue network	3.879	$\beta_1 = .029, \beta_2 = 0, \beta_3 = .030, \beta_4 = .043, \beta_b = 0, c = 2.220$
Context	2.927	$a_1 = .995, a_2 = .839, a_3 = 1.266\ a_4 = 1.430, b = .0078, B = .629$
Exemplar-based network (ALCOVE)	2.849	$\alpha_1 = 1.033, \alpha_2 = 32.681, \alpha_3 = 2.582, \alpha_4 = 1.398, \beta = .152, c = 1.503$
Configural-cue (Version 1)	2.861	$\beta_1 = 0, \beta_2 = .031, \beta_3 = 0, \beta_4 = 0, \beta_b = 0, c = 1.424$
Configural-cue (Version 2)	3.003	$\beta = .086, \gamma_1 = .308, \gamma_2 = .285, \gamma_3 = .361, \gamma_4 = .730, \beta_b = 0, c = 1.325$

Note. SSD = sum of squared deviations. ALCOVE = attentional learning covering map.

model and ALCOVE. Future research will be needed to sharply contrast the predictions of the configural-cue model with those of the context model and ALCOVE.

General Discussion

The main theme of this research was to compare and contrast the component-cue model, a learning version of the context model, and an exemplar-based network model (a version of Kruschke's [1992] ALCOVE model) on their ability to predict category learning and transfer. The exemplar-based network incorporates the same exemplar-based category representation, similarity rules, and selective attention processes that are assumed in the context model but combines them with the error-driven learning rules that are assumed in adaptive networks such as the component-cue model.

In Experiment 1 we conducted a partial replication and extension of the probabilistic classification learning paradigm tested previously by Gluck and Bower (1988a) and Estes et al. (1989). All models provided equally good accounts of the learning data, but the component-cue model and the exemplar-based network outperformed the context model in predicting transfer performance. The network models were able to characterize a form of base-rate neglect that was observed during transfer, but the context model was not. The superiority of the network models in this domain was attributed to their use of an error-driven, interactive learning rule.

In Experiment 2 we conducted a partial replication and extension of a category learning paradigm used extensively by

Table 8
Models Fitted Simultaneously to the Learning and Transfer Data in Experiment 2

Model	Learning SSD	Transfer SSD	Total SSD	Parameters
Baseline				
Component-cue network	4.629	2.344	6.955	$\beta = .035, \beta_b = .001, c = 1.828$
Context	2.986	.496	3.481	$a = 1.258, b = .0078, B = 1.130$
Exemplar-based network (ALCOVE)	3.067	.489	3.556	$\kappa = 6.018, \beta = .119, c = 1.492$
Configural-cue	3.106	.804	3.902	$\beta = .0092, \beta_b = 0, c = 1.399$
Elaborated				
Component-cue network	4.093	2.366	6.442	$\beta_1 = .030, \beta_2 = .001, \beta_3 = .028, \beta_4 = .042, \beta_b = 0, c = 1.903$
Context	2.943	.466	3.409	$a_1 = 1.051, a_2 = 1.105, a_3 = 1.069\ a_4 = 1.252, b = .0096, B = .518$
Exemplar-based network (ALCOVE)	2.981	.510	3.491	$\alpha_1 = 1.355, \alpha_2 = 1.873, \alpha_3 = 1.456, \alpha_4 = 1.488, \beta = .127, c = 1.499$
Configural-cue (Version 1)	3.122	.580	3.691	$\beta_1 = 0, \beta_2 = .0106, \beta_3 = .0012, \beta_4 = .0248, \beta_b = 0, c = 1.614$
Configural-cue (Version 2)	3.127	.591	3.708	$\beta = .0017, \gamma_1 = 1.645, \gamma_2 = 3.362, \gamma_3 = 1.671, \gamma_4 = 1.219, \beta_b = 0, c = 1.518$

Note. SSD = sum of squared deviations. ALCOVE = attentional learning covering map.

Medin and his associates. This paradigm sharply contrasts the predictions of exemplar models (using nonlinear similarity rules) with those of prototype models. The paradigm was very effective for demonstrating limitations of the component-cue model, which is in essence a multiplicative-similarity prototype model. A variety of prototype-enhancement phenomena that were predicted by the component-cue model were not observed in Experiment 2, and the quantitative fits of the model were extremely poor. By contrast, both the context model and the exemplar-based network model performed admirably by predicting accurately the average classification probabilities for individual patterns, the trial-by-trial sequence of learning data, and the evolution of transfer performance over blocks.

Taken together, the results of our experiments suggest that the most promising of the three models is the exemplar-based network model. As demonstrated in Experiment 1, its error-driven, interactive learning rule gives it an advantage over the context model by allowing it to predict base-rate neglect phenomena in probabilistic classification paradigms.[8] Moreover, as demonstrated in Experiment 2, its exemplar-based category representation gives it an advantage over the component-cue model by allowing it to predict exemplar-based generalization processes.

There are several important directions for future research. One such direction is to test systematically whether the precise quantitative predictions achieved by the context model in modeling categorization, identification, and recognition data in previous research (see Nosofsky, in press-b, for a review) can be matched by the exemplar-based network (ALCOVE). Although the models are closely related, there is no guarantee that modifying the context model by incorporating error-driven learning will not adversely affect its previous successes.

Another important direction is to conduct careful comparisons between ALCOVE and Gluck and Bower's (1988b) configural-cue model. In our view there are three main distinctions between the models as they are currently articulated. First, although both models incorporate exemplar nodes that become associated to categories by an error-driven learning rule, the configural-cue model assumes that all lower order configurations of features also develop associations. Second, activation of nodes in the configural-cue model is all-or-none, whereas activation of nodes in ALCOVE is proportional to the similarity between a node and an input pattern. Third, ALCOVE has mechanisms for attentional learning, in which certain dimensions are weighted more heavily than others in calculating similarity. Selective attention to dimensions is not part of Gluck and Bower's (1988b) configural-cue model. These investigators proposed that attentional phenomena might emerge from their system but did not build in attentional mechanisms.

We view the aspect of the configural-cue model in which nodes are activated in an all-or-none, discrete fashion as a major shortcoming of that model. Imagine that the network has been trained on a set of 1-inch and 3-inch objects and that during transfer a 2.9-inch object is presented. In its current form, the model has no way of incorporating information that the 2.9-inch object is more similar to the 3-inch objects than to the 1-inch objects, thus it would be unable to predict appropriate generalization behavior. Such similarity-based generalization is a fundamental assumption in ALCOVE.

We also believe that the lack of attentional learning is another shortcoming of the configural-cue model. Although attentional learning was not critical for explaining performance in our experiments, Nosofsky (1984, 1987, 1989) and Kruschke (1990a, 1992) have demonstrated its importance in other settings. Indeed, Gluck and Chow (1989) have discussed the need to build some type of attention-learning mechanism into the configural-cue model and have conducted preliminary work along these lines.

Once similarity-based activation of nodes and attentional learning are added to the configural-cue model, it will differ from ALCOVE mainly in its assumption that all lower order configurations of features are part of the category representation. Whether these lower order nodes improve or detract from the model's predictions then becomes an interesting question to pursue. At least in Experiment 2 of our study, we obtained preliminary evidence that the lower order nodes detracted from the model's predictions, but we have no idea how general this result might be.

[8] Of course, there are other ways of modifying the context model to incorporate error-driven learning (Medin & Edelson, 1988), but rigorous quantitative formulations along these lines have not yet been proposed.

References

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3–11.

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

Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115,* 155–174.

Estes, W. K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language, 27,* 196–212.

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

Garner, W. R. (1974). *The processing of information and structure.* New York: Wiley.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32–41.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2,* 50–55.

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

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

Gluck, M. A., Bower, G. H., & Hee, M. R. (1989). A configural-cue network model of animal and human associative learning. *Pro-*

ceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, MI.

Gluck, M. A., & Chow, W. (1989). *Dynamic stimulus-specific learning rates and the representation of dimensionalized stimulus structures.* Unpublished manuscript.

Golden, R. M., & Rumelhart, D. E. (1989). *A relationship between the context model of classification and a connectionist learning rule.* Unpublished manuscript.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Hurwitz, J. B. (1990). *A hidden-pattern unit network model of category learning.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Kruschke, J. K. (1990a). *A connectionist model of category learning.* Doctoral dissertation, University of California, Berkeley. Available from University Microfilms International.

Kruschke, J. K. (1990b). *ALCOVE: A connectionist model of category learning* (Research Report #19). Cognitive Science Program, Indiana University.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97,* 225–252.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol 24, pp. 109–165). New York: Academic Press.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 3,* 333–352.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

Medin, D. L., & Florian, J. E. (in press). Abstraction and selective coding in exemplar-based models of categorization. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2). Hillsdale, NJ: Erlbaum.

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

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 241–253.

Myung, I. J., & Busemeyer, J. R. (in press). Measurement free tests of a general state space model of prototype learning. *Journal of Mathematical Psychology.*

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

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics, 38,* 415–432.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–109.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279–290.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34,* 393–418.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

Nosofsky, R. M. (in press-a). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M. (in press-b). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition.*

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1991). *Comparisons between adaptive network and exemplar models of classification learning* (Research Report No. 35). Bloomington: Indiana University, Cognitive Science Program.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Freidman (Eds.), *Handbook of perception* (Vol. 2, pp. 128–141). New York: Academic Press.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

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

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

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55,* 509–523.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

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

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

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

(Appendix follows on next page)

Appendix

Tests of Alternative Configural-Cue Models

In this Appendix we consider the standard configural-cue model on its ability to predict the classification data in Experiment 1 and also consider some alternative versions of the model than were discussed in the text. These alternative models are tested on both the Experiment 1 and Experiment 2 data.

As discussed in the text, the standard configural-cue model is the same in all respects as the component-cue model, except that the input nodes code all configurations of features instead of only the single features. Similar to the component-cue model, the configural-cue model has three free parameters, as follows: learning rates β and β_b for the feature nodes and bias node, respectively; and the logistic scale parameter c for transforming the output-node activations to response probabilities.

We also tested two elaborated versions of the standard configural-cue model. In one version, separate learning rates were allowed for nodes of differing dimensionality (i.e., the single, double, triple, and exemplar nodes). A special case of this model that is of interest arises when the learning rates on the triple nodes and the exemplar nodes are set to zero, leaving just the single and double nodes. We refer to this model as the *doublet* model. Gluck, Bower, and Hee (1989) have reported some successes with such a model.

In a second elaborated version, separate salience parameters (γ_m) were defined for each dimension m. The learning rate on node k (β_k) was then given by

$$\beta_k = \beta\left(\prod \gamma_m^{\delta(m,k)}\right),$$

where $\delta(m, k) = 1$ if dimension m is part of node k's configuration, and $\delta(m, k) = 0$ otherwise. Note that this second elaborated version allows for differential learning rates that are sensitive to both individual dimension salience and the overall dimensionality of nodes. In general, when the γ_ms are greater than one, higher order nodes receive higher learning rates than lower order nodes, whereas when the γ_ms are less than one, the reverse occurs.

Gluck (personal communication, April 1991) suggested that the logistic transformation may be inappropriate when used with the configural-cue model. The basic argument is that, because it codes all configurations of features, the configural-cue model already embodies a nonlinear relation between the number of feature matches and similarity between patterns (see Gluck, 1991). Use of the logistic response rule on top of this nonlinear similarity mapping may hurt the model's predictions. As an alternative, Gluck (personal communication, April 1991) suggested use of a raw-output response rule. In this model, the teaching signals at the output nodes are set at 0 and 1 (instead of −1 and 1). To predict the probability of a Category A response, one takes the Category A output and divides by the sum of the outputs to Category A and Category B, with occasional negative outputs truncated to zero. We consider this assumption of truncating negative outputs to be inelegant, but test it here in an attempt to be as fair as possible to the configural-cue model. (Note that the standard model with the logistic transformation does not run into this problem.) We refer to this alternative version of the configural-cue model as the raw-output model. We also fitted elaborated versions of the raw-output model that were analogous to our elaborated versions of the standard model.

Experiment 1 Analyses

Table A1 (Experiment 1) shows the fits of the different versions of the configural-cue model to the learning and transfer data. The

Table A1
Fits of Versions of the Configural-Cue Model to the Learning and Transfer Data

Model	Transfer SSD	Condition of learning SSD		No. of free parameters
		Categorization	Estimation	
		Experiment 1		
Logistic output				
Standard	.947	1.900	2.116	3
Elaborated 1	.482	1.831	2.067	6
Doublet	.482	1.831	2.067	4
Elaborated 2	.213	1.661	1.777	7
Raw output				
Standard	.604	3.506	3.623	2
Elaborated 1	.569	3.478	3.615	5
Doublet	.569	3.502	3.661	3
Elaborated 2	.391	3.054	2.998	6
		Experiment 2		
Logistic output				
Standard	.593	3.065		3
Elaborated 1	.342	2.861		6
Doublet	.688	2.861		4
Elaborated 2	.415	3.003		7
Raw output				
Standard	.688	3.574		2
Elaborated 1	.346	3.518		5
Doublet	.891	3.564		3
Elaborated 2	.355	3.376		6

Note. SSD = sum of squared deviations.

number of free parameters allowed for each version of the model is also shown. (The values of the best-fitting parameters are available on request.)

The raw-output model performs far worse than the logistic-output model at predicting the learning data in both the categorization and estimation conditions and, in general, performs roughly the same as the logistic-output model at predicting the transfer data. The raw-output model's difficulties with the learning data are probably due to its truncating of negative outputs to zero. For example, early in the learning sequence, suppose the output to Category A is very slightly positive and the output to Category B is very slightly negative. Despite the small magnitudes of the outputs, the raw-output model nevertheless would predict that the item would be classified in Category A with a probability of one. Because of these difficulties in fitting the learning data, and because the models perform roughly the same on the transfer data, we focus the remainder of our discussion on the logistic-output configural-cue model.

As shown in Table A1 (Experiment 1) the standard configural-cue network fits the transfer data far worse (sum of squared deviations [SSD] = .947) than does the component-cue network, ALCOVE, and even the context model (see Table 3). By allowing the learning rate on nodes of differing dimensionality to be free parameters, the elaborated configural-cue model is able to match the fit of the component-cue network (SSD = .482; compare with Table 3). Indeed, for this model, the best-fitting parameters on all higher order nodes were essentially zero, thus the configural-cue model essentially becomes the component-cue network. (The same is true for the doublet model.) Finally, the elaborated version in which each individual dimension was allowed a separate salience parameter performs the best (SSD = .213). The fit of this model is basically the same as for the elaborated versions of the component-cue network and ALCOVE (see Table 3).

In summary, the standard configural-cue model performs worse than the component-cue network and ALCOVE at fitting the transfer data but, by elaborating the model with free parameters, the multiplicative-salience version performs as well as the elaborated component-cue network and ALCOVE. All models perform roughly the same on the learning data, except for the raw-output configural-cue models, which perform markedly worse than the other models.

Experiment 2 Analyses

Table A1 (Experiment 2) shows the fits (total SSD) of the different versions of the configural-cue model to the learning and transfer data. As discussed in the text, the standard configural-cue model performs worse (SSD = .593) than ALCOVE and the context model at predicting the transfer data (compare with Table 6). Furthermore, using the raw-output function does not help the standard model (SSD = .688). The doublet model, in which learning rate on the triple nodes and exemplar nodes is set at zero, also performs quite poorly, regardless of whether a logistic-output or raw-output function is used (SSDs = .688 and .891, respectively). As explained previously, the reason for the poor performance of these versions of the model is that the single nodes tend to predict strong prototypicality effects that were not observed in our Experiment 2 data.

The only versions of the model that perform well on both the learning and transfer data are the elaborated versions with a logistic-output function, in which free parameters are allowed for describing learning rates on different nodes. In both elaborated models, the free parameters took on values that deleted the single nodes from the learning process and concentrated most of the learning on the higher order nodes, especially the exemplar nodes.

Received December 28, 1990
Revision received July 8, 1991
Accepted August 15, 1991 ■