Psychological Review
1978, Vol. 85, No. 3, 207–238

Context Theory of Classification Learning

Douglas L. Medin
Rockefeller University

Marguerite M. Schaffer
Barnard College

Most theories dealing with ill-defined concepts assume that performance is based on category level information or a mixture of category level and specific item information. A context theory of classification is described in which judgments are assumed to derive exclusively from stored exemplar information. The main idea is that a probe item acts as a retrieval cue to access information associated with stimuli similar to the probe. The predictions of the context theory are contrasted with those of a class of theories (including prototype theory) that assume that the information entering into judgments can be derived from an additive combination of information from component cue dimensions. Across four experiments using both geometric forms and schematic faces as stimuli, the context theory consistently gave a better account of the data. The relation of the context theory to other theories and phenomena associated with ill-defined concepts is discussed in detail.

One of the major components of cognitive behavior concerns abstracting rules and forming concepts. Our entire system of naming objects and events, talking about them, and interacting with them presupposes the ability to group experiences into appropriate classes. Young children learn to tell the difference between dogs and cats, between clocks and fans, and between stars and street lights. Since few concepts are formally taught, the evolution of concepts from experience with exemplars must be a fundamental learning phenomenon. The focus of the present article is to explore how such conceptual achievements emerge from individual instances.

Structure of Concepts

An early step in analyzing task demands involved in conceptual behavior is to ask how individual instances or exemplars are related to the superordinate concept. Although there is general agreement that natural categories are structured so that exemplars within a category are more similar to one another than to exemplars from alternative categories, there is disagreement concerning the rigidity of this structure. One extreme view is that all natural concepts are characterized by simple sets of defining features that are singly necessary and jointly sufficient to determine category membership (Katz & Postal, 1964). Each exemplar of the concept must possess these defining features, and therefore, each exemplar is equally representative of the concept. Concepts containing singly necessary and jointly sufficient defining features are said to be well-defined concepts.

A contrasting point of view is that most natural concepts are not well-defined but rather are based on relationships that are only generally true. Individual exemplars may vary in the number of characteristic features they possess, and consequently, some exemplars may be more representative or more typical of a concept than others. For example, cows may be better exemplars of the concept *mammal* than are whales. Instances are neither arbitrarily associated with categories nor strictly linked by defining features, but rather in-

This research was supported by United States Public Health Grant MH16100 and by Grant MH 23878 from the National Institute of Mental Health.

We wish to acknowledge the support of Edith Skaar, who assisted in all phases of this research, and the patience of Mark Altom, Lee Brooks, Donald Robbins, and Edward E. Smith, who read earlier drafts of this article and provided helpful suggestions.

Requests for reprints should be sent to Douglas L. Medin, Box 298, Rockefeller University, New York, New York 10021.

Copyright 1978 by the American Psychological Association, Inc. 0033-295X/78/8503-0207$00.75

stances reflect more nearly a "family resemblance" structure (Rosch & Mervis, 1975).

These ideas concerning the structure of categories influence the way instances are set up in laboratory studies of artificial concepts. Most early work with concepts used well-defined concepts and focused on such issues as the relative difficulty of acquiring different rules, strategies for formulating and testing alternative hypotheses, and the transfer of behavior to new stimulus sets (e.g., Bourne, 1970; Levine, 1975; Trabasso & Bower, 1968). While this approach has accumulated considerable information about processes underlying hypothesis selection and rule learning where rules are well-defined, little information exists to show how these models might be applied to a variety of other situations where rules and concepts are not so precisely defined. Therefore, substantial reason exists to examine more closely the case in which the concepts and classifications acquired in everyday experience do not conform to well-defined rules.

On the basis of an extensive series of experiments, Rosch and her associates (Rosch, 1973, 1975a, 1975b, 1975c; Rosch & Mervis, 1975; Rosch et al., 1976) have argued that most natural categories do not have well-defined rules or fixed boundaries separating alternative categories. Rather, members vary in the degree to which they are judged to be good examples (typical) of the category, and many natural concepts cannot be defined by a single set of critical features. In addition, subjects appear to use nonnecessary features in making category judgments. Smith, Shoben, and Rips (1974) found that the items judged to be typical of a category possess features that are characteristic of the class but not necessary for category definition. For example, *robin* is a typical member of the category *bird* and has the characteristic feature that it flies, but not all birds fly (e.g., penguins). In a reaction time task, Smith et al. observed that subjects required less time to verify the category membership of the more typical items in a category. Characteristic features and not just defining features appear to be involved in these category judgments.

If many natural categories have a loosely defined structure, how do people acquire and use this structure? Posner and Keele (1968) proposed that based on experience with exemplars, people form an impression of the central tendency of a category and that categorical judgments come to be based on this central tendency, or prototype. While there is not universal agreement that prototype formation underlies conceptual learning in this domain, the increasing evidence that natural categories and concepts are not well-defined has amplified the interest in developing theories of conceptual behavior appropriate to rules with exceptions.

The present article takes the perspective of aiming to see if recent theoretical developments arising in the domain of discrimination learning might be profitably applied to classification learning. In the case of well-defined rules for stimulus classification, there are some striking parallels between paradigm and theory in discrimination learning and concept identification. For example, the simple affirmative concept "red in one pile, green in the other pile" corresponds to a simultaneous discrimination learning task, where red is correct and green is incorrect. Likewise, hypothesis-testing theories for concept identification tasks (e.g., Trabasso & Bower, 1968) are closely mirrored by theories of selective attention in discrimination learning (Medin, 1976; Sutherland & Mackintosh, 1971).

Is there any basis for expecting useful interaction between the domains of discrimination learning and concept learning for ill-defined rules? We shall argue that there is. Not all discrimination learning problems map onto simple affirmative rules. For example, in a successive brightness discrimination problem, the solution might be "If the choice stimuli are white, go right; if black, go left." The various stimulus components, that is, black, white, left, and right, each are associated with reward half the time, and the problem could not be solved on the basis of associations to these independent stimulus components. Indeed, Spence's (1936) theory of discrimination learning assumed independence of components, and it was unable to predict that successive discrimination problems could be mastered. Other discrimination learning theories have been proposed that can account for relationships between simultaneous and successive discrimination learning, and the present article

attempts to demonstrate the applicability of one such theory (Medin, 1975) to learning and classification involving ill-defined categories. Before discussing this theory, however, we consider two phenomena that have been adduced in support of prototype theory and that directly motivated our theoretical efforts.

Some Evidence Related to Classification Involving Ill-defined Concepts

In a typical study assessing the learning of ill-defined concepts, subjects learn to sort a set of instances into two or more categories and then are given transfer tests with new stimuli, including a pattern representing the central tendency, or prototype. An alternative to the procedure of presenting two or more contrasting categories is simply to present subjects with instances of a single concept and then to give a new–old recognition test for new and old instances. Stimuli range from schematic faces, geometric forms, and dot patterns to letter sequences and biographical descriptions. A major theoretical view is that as a function of experience with exemplars of a category, subjects abstract out the central tendency of the category. This summary representation, or prototype, is assumed to provide the basis for classification performance. The closer an exemplar is to its category prototype and the farther it is from the prototypes associated with alternative categories, the greater the likelihood that it will be appropriately classified.

Two main results arising from this literature seem to provide cogent evidence for prototype formation:

1. Subjects classify prototypic patterns they have never seen before virtually as fast and accurately as they classify old training patterns. In addition, these two types of patterns are classified better than other new exemplars (e.g., Homa & Chambliss, 1975; Homa & Vosburgh, 1976; Peterson, Meagher, Chait, & Gillie, 1973; Posner & Keele, 1968, 1970; Strange, Keeney, Kessel, & Jenkins, 1970). Transfer performance is well-predicted by distance of a pattern from the prototype and not by the frequency with which individual features of patterns appeared during training (e.g., Franks & Bransford, 1971; Lasky, 1974; Posnansky & Neumann, 1976).

It is generally proposed that two factors determine classification performance in these situations (e.g., Posner & Keele, 1968, 1970). One is specific item information, which leads to old training patterns being classified better than new patterns; the other is abstraction, which gives rise to performance being a function of the distance of a pattern from the central tendency, or prototype.

2. The second major finding concerns differential retention of old and new patterns. When delays on the order of several days are inserted between learning and transfer tests, significantly greater forgetting is observed for the old training stimuli than for the prototype and other new patterns (Goldman & Homa, 1977; Homa et al., 1973; Homa & Vosburgh, 1976; Posner & Keele, 1970; Strange et al., 1970). These results are consistent with the idea that judgments are based on a mixture of specific item and category level information and that the specific item information is forgotten more rapidly than the abstract, category level information. According to this view, as retention interval increases, judgments are increasingly likely to be based on the prototype and less likely to be based on specific exemplar information.

Previous theoretical explanations of the above two phenomena have relied on positing both specific item and category level information. Yet, we noted that at least one discrimination learning theory (Medin, 1975) might account for these results and interactions simply in terms of the similarity or confusability of the learned exemplars. Therefore, it is of interest to see how well this theory can account for classification without assuming that category level information influences performance. In the remainder of this article, we present a context theory of classification based on Medin's (1975) context theory of discrimination learning, describe four new experiments that contrast the predictions of the context model with predictions derived from a large class of classification models, and finally discuss more generally the relationship between the context theory and classification behavior.

Context Theory for Classification

The general idea of the context model is that classification judgments are based on the

retrieval of stored exemplar information. Specifically, we assume that a probe stimulus functions as a retrieval cue to access information stored with stimuli similar to the probe. This mechanism is, in a sense, a device for reasoning by analogy inasmuch as classification of new stimuli is based on stored information concerning old exemplars, but one should also note its similarity to contemporary memory models. For example, Ratcliff (Note 1) has recently proposed a model for recognition memory that relies on a resonance metaphor, whereby the set of stored items to be searched is evoked on the basis of similarity of the information provided by the probe to the information in the memory items.

Although we shall propose that classifications derive from exemplar information, we do not assume that the storage and retrievability of this exemplar information is necessarily veridical. If subjects are using strategies and hypotheses during learning, the exemplar information may be incomplete and the salience of information from alternative dimensions may differ considerably. A person focusing on the hypothesis that red cards belong in Category A and green cards in Category B might store and be able to retrieve little information concerning, say, the form of the stimuli. The context model attempts to represent the effect of strategies and hypotheses in terms of the ease of storage and retrieval of information associated with the component stimulus dimensions.

Notation

In discussing the context theory and alternative theories, it is useful to work with an abstract notation. Consider a situation where the stimuli to be classified comprise binary values on each of four dimensions (e.g., color [red or blue], form [triangle or circle], size [large or small], and number [one or two]). By assigning the number 1 to one value on a dimension and the number 0 to the other, each stimulus can be described in terms of a simple binary code. For example, if the dimensions are ordered as color, form, size, and number, the notation 1001 might refer to a single small red circle, while 0110 would refer to a stimulus comprised of two large blue triangles. This binary code is also useful for noting similarity relationships. Thus, the two stimuli 1111 and 1101 differ only in size, while 0101 and 1101 differ only in color.

Before proceeding to the particular assumptions of the context model, it might prove helpful to consider a specific example. Suppose two A patterns, A_1 and A_2 (A_1 = 1110 and A_2 = 1010), and two B patterns, B_1 and B_2 (B_1 = 0001 and B_2 = 1100), have been presented a few times during a classification learning task and that now a new stimulus 1101 is presented. Suppose further that the person in the experiment has selectively attended to the first two dimensions, so that less information has been stored or is retrievable concerning the third and fourth dimensions. The subject's representation of exemplar information may be something like this:

$$\begin{array}{ll} 111?\text{–}A(A_1) & 10?0\text{–}A(A_2) \\ 00?1\text{–}B(B_1) & 110?\text{–}B(B_2), \end{array}$$

where the question marks indicate that information that would differentiate value 1 and value 0 on that dimension either has not been stored or cannot be accessed. When the probe 1101 is presented, the most likely event is that the representation associated with the B_2 exemplar (110?) will be retrieved, and the probe will be classified as a B. The next most likely event is that exemplar A_1 will be retrieved, since the probe differs from the representation of A_1 only on the third dimension. In this case, the probe would be classified as an A. The greater the similarity of a stored exemplar to a probe, the more likely it is that the probe will retrieve the information associated with that exemplar. In the following paragraphs, we provide a general rationale for our ideas concerning similarity and then develop the specific assumptions of the context model.

Interactive Similarity

Imagine that a blue circle is presented in some experimental context and that some event (e.g., the classification assignment) occurs. It is not assumed that individual stimulus components get directly and independently associated with the event. Rather, we propose that information concerning the cue, the con-

text, and the event are stored together in memory and that both cue and context must be activated simultaneously in order to retrieve information about the event. A change in either the cue or the context can impair the accessibility of information associated with both.

This idea is depicted in Figure 1. R(cue), R(context), and R(event) refer to the memory representation of the cue, context, and event, respectively. It is further assumed that a particular stimulus component serves a cue function and acts as context for other cues. In our example, blue is part of the context in which circle appears. As a result, transfer or generalization along the form dimension will not be independent of color value. This nonindependence represents an important constraint on the accessibility of stored information and provides the basis for the experimental contrasts to be considered later.

This formulation is closely related to the assumptions of the Estes hierarchical association model (Estes, 1972, 1973, 1976). The cue–context node corresponds to what Estes calls a "control element," and we use it to denote the assumptions (a) that neither cue nor context is directly associated with an event or outcome and (b) that inputs from both cue and context are needed to activate the node and provide access to the representation of an event. The latter assumption implies that the effect of cue changes and context changes combine in an interactive manner.

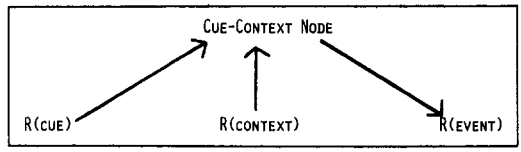

Figure 1. Illustration of factors proposed to determine the accessibility of information associated with a cue presented in a particular context. (*R* refers to the memory representation of the cue, context, or event.)

Specific Assumptions

1. Category judgments are based on the retrieval of specific item information; no categorical information is assumed to enter into the judgments independently of specific item information. While it is assumed that categorical information does not influence judgments, this is not the same as assuming that category level information does not exist. In the case of natural categories, information on the level of categories is often explicitly presented. Our proposal simply is that judgments in classification tasks are based on retrieval of exemplar information rather than on category level information.

2. The probability of classifying exemplar i into category j is an increasing function of the similarity of exemplar i to stored category j exemplars and a decreasing function of the similarity of exemplar i to stored exemplars associated with alternative categories. Specifically, it is assumed that the evidence favoring a category j response to probe i is equal to the sum of the similarities of probe i to the stored j exemplars divided by the sum of the similarities of probe i to all stored exemplars. For purposes of the present article, we assume that the probability of a j response is equal to the evidence favoring a j classification. The mechanism by which these similarities operate is detailed in the next assumption.

3. Probe or test stimuli act as retrieval cues to access information associated with stimuli similar to the probe. Which stimuli will be retrieved depends on the overall similarity of the stored exemplars to the probe stimulus. Instead of proposing that subjects compute the similarity of a probe to all of the training patterns, we assume that the retrieval rules act to determine which patterns are likely to be accessed. In fact, later on we shall consider the possibility that judgments are based solely on the first pattern retrieved.

4. The similarity of two cues along a dimension can be represented by a similarity parameter whose value can range between 0 and 1, with 1 representing maximum similarity. For example, the similarity of a yellow circle and a blue triangle along the color dimension would be represented by a parameter c for color similarity, while form similarity would be represented by a parameter f. The parameter c for color would be larger if the two colors were yellow and orange than if the two colors were yellow and blue, since presumably yellow is more similar to orange than to blue.

5. The various cue dimensions comprising stimuli in some context are combined in an

interactive, specifically multiplicative, manner to determine the overall similarity of two stimuli. This means that the overall similarity of a yellow circle and a blue triangle would be equal to cf. If the form differences were very salient (i.e., $f \cong 0$), then variations in color similarity would not alter performance, since $cf \cong 0$ regardless of the value of c.

Although the rationale for the multiplicative rule grows out of attempts to represent the effects of similarity and context in other situations (e.g., Estes, 1973; Medin, 1975), it fits well with certain intuitions concerning natural concepts. For example, although mannequins may resemble human beings along a variety of dimensions, we are not tempted to judge that they are, in fact, human beings. In a multiplicative combination rule, the effects of various dimensions of similarity can be effectively overcome by a single dimension of difference (e.g., animacy). If judgments were based solely on a summing of evidence, it would be more awkward to represent the difference between mannequins and human beings. In short, the interactive rule has the potential to represent the effects of necessary features without the theory committing itself to the idea that category membership is defined in terms of singly necessary and jointly sufficient features.

The multiplicative rule implies that a pattern will be classified more efficiently if it is highly similar to one pattern (differing in only one dimension) and has low similarity to a second (differing in three dimensions) than if it has medium similarity (differing in two dimensions to two patterns in its category. If all dimensions were equally salient, then in the first case, the net similarity would be $s + s^3$ (where s is the similarity parameter for a difference along a given dimension); and in the second, the net similarity would be $2s^2$. Since s is defined to be between 0 and 1, the first term is larger than the second, except for the uninteresting cases where s is 0 or where s is 1.

6. Selective attention can be represented by changes in the salience or similarity parameter for dimensions. That is, the similarity parameter of two cues along a dimension is less when that dimension is attended than when it is not attended.

This assumption is designed to capture the consequences of active hypothesis testing. For example, if subjects were trying out the possibility that all red stimuli belong to Category A and all green stimuli to Category B, they might code much less information about other attributes such as size or form than otherwise. As a result, the effective similarity of two size or two form cues might be greater than usual, and the effective similarity of red and green would be expected to be less than otherwise. For tests that can be solved by attending to a single dimension, subjects may have only minimal information to distinguish the individual exemplars (see Bourne & O'Banion, 1969; Calfee, 1969).

7. The last general assumption is that retention loss can be represented by changes (increases) in the similarity parameters. We assume that retention loss corresponds to a loss of distinctiveness, which we represent as increased similarity, regardless of whether it arises from forgetting of hypotheses or strategies involving selective attention or from decreased availability of the exemplar information independent of strategies.

No special assumptions are made to distinguish old and new patterns, and later on we will indicate how differential forgetting is handled in terms of changes in the similarity parameters. We propose that the ease of learning to classify a training pattern into Category A and the likelihood of classifying a new pattern into Category A increases with the similarity of the pattern to the stored exemplars in Category A and decreases with the similarity of an item to the exemplars in Category B. One shortcoming of the context model as so far developed is that specific assumptions concerning learning and storage of exemplar information have not been spelled out. Predictions concerning errors during learning will be based on the qualitative evidence derived from the similarity of an exemplar to other exemplars of its category versus its similarity to exemplars of alternative categories.

Application to an Example

In developing the predictions of the context theory as well as those alternative classification theories, it is useful to work with a con-

crete example. Consider a situation where six stimuli are to be presented, three assigned to Category A and three to Category B. Suppose further that the stimuli comprise binary values on the dimensions of color, form, size, and position. Figure 2 represents this situation in terms of our abstract notation. If 1 represents red and 0 blue, then red figures are more often associated with Category A than Category B. One can also see that Stimulus 1 differs from Stimulus 2 only on position but differs from Stimulus 3 in color, form, and size. The categories are not well-defined because there is no set of singly necessary and jointly sufficient features to define category membership. In fact, no simple feature is even necessary for category membership. Note that for any dimension, two of the three Category A members have value 1 and two of the three Category B members have value 0. We will now use this notation and write prediction equations based on our assumptions concerning similarity and retrieval.

The similarity parameters for the dimensions of color, form, size, and position are represented as parameters c, f, s, and p when two comparison stimuli differ in their value on the dimension in question, and the similarity parameter is set at 1 when the two values are identical. Using this notation for Stimulus Pattern 1, we can represent the evidence favoring a Category A assignment (E_A) in terms of the likelihood that the probe will retrieve a pattern associated with Category A as

$$E_{A,1}=[(1\cdot1\cdot1\cdot1)+(1\cdot1\cdot1\cdot p)+(c\cdot f\cdot s\cdot 1)] \div[(1\cdot1\cdot1\cdot1)+(1\cdot1\cdot1\cdot p)+(c\cdot f\cdot s\cdot 1)+(c\cdot f\cdot s\cdot p)+(c\cdot f\cdot 1\cdot 1)+(1\cdot 1\cdot s\cdot p)]. \quad (1)$$

Multiplying out the terms in parenthesis, we can write the more simple form

$$E_{A,1} = \frac{1 + p + cfs}{1 + p + cfs + cfsp + cf + sp}. \quad (2)$$

The numerator of Equation 2 arises from the fact that Stimulus 1 is identical to itself; differs only in position (p) from Stimulus 2; and differs in color, form, and size (cfs) from Stimulus 3. The second three terms of the denominator represent the overall similarity of Stimulus 1 to Stimulus 4, Stimulus 5, and Stimulus 6, respectively.

Category A					Category B				
Stimulus	Pattern				Stimulus	Pattern			
	C	F	S	P		C	F	S	P
1	1	1	1	1	4	0	0	0	0
2	1	1	1	0	5	0	0	1	1
3	0	0	0	1	6	1	1	0	0

Transfer				
Stimulus	Pattern			
	C	F	S	P
7	0	1	0	1

Figure 2. Abstract notation for representing stimuli with binary values along four dimensions. (C, F, S, and P stand for color, form, size, and position, respectively.)

In a directly analogous manner, one can show that for Stimulus 4 in Figure 2,

$$E_{B,4} = \frac{1 + sp + cf}{1 + sp + cf + cfsp + p + cfs}. \quad (3)$$

Comparing Equations 2 and 3, one may note that if $c = f = s = p = x$ ($x \neq 0$ or 1), $E_{A,1}$ is larger than $E_{B,4}$. The denominators of the two equations are identical; and subtracting the numerator of Equation 3 from that of Equation 2, we get $x - 2x^2 + x^3$, which is positive because x lies between 0 and 1. Based on the general proposal that ease of learning and accuracy of classification increase as the evidence favoring the category increases, Pattern 1 should be easier to learn and classify than Pattern 4.

On a qualitative level, one may note that the reason that Stimulus 1 should be easier to learn and classify than Stimulus 4 is that Stimulus 1 is highly similar to one other pattern (Stimulus 2), and this pattern is associated with the same category; while Stimulus 4 is highly similar to one other pattern (Stimulus 3), but this pattern is associated with the opposite category. Therefore, when Stimulus 4 is presented, the representation and category assignment associated with Stimulus 3 might well be activated and produce misclassifications.

Because of its multiplicative combination rule for deriving overall similarity, the context model implies that performance will be primarily affected by stored exemplars that are highly similar to the item in question. This relationship is assumed to hold for transfer as

well as original learning. Usually classification models make qualitative predictions based on the evidence favoring one category versus alternative categories. We shall follow this practice in comparing the context model with alternative models, but at the same time, we shall attempt to make quantitative predictions by using a simple rule relating the weights to classification probabilities. As a first attempt at quantitative predictions, we will let the probability that an item will be classified as an A on transfer tests be equal to the weight favoring Category A. That is,

$$P_{A,i} = E_{A,i}. \tag{4}$$

For example, for New Stimulus 7 in Figure 2, we would have

$$P_{A,7} = \frac{cs + csp + f}{cs + csp + f + fp + fs + cp}, \tag{5}$$

where f, s, c, and p are the respective similarity parameters for the dimensions of form, size, color, and position. If all four parameters were equal to the value .30, then according to Equation 5, New Pattern 7 should be classified as an A 61% of the time.

In brief, the context model assumes that classification is based on the retrieval of stored exemplar information, that this retrieval process is directly a function of the similarity of the probe item to the stored exemplars, and that similarity derives from an interactive combination of component cue dimensions. We consider now how these ideas might account for excellent performance on prototype stimuli and differential retention.

Performance on Prototypic Stimuli and Differential Retention

Performance on prototype patterns frequently equals or exceeds performance on old training patterns. This result can be predicted by the context model because the prototypic pattern almost always is the pattern having the greatest number of highly similar category exemplars or training patterns. Not only that but also because of the ways in which categories have usually been constructed (Reed, 1972), the prototype is unlikely to be highly similar to any exemplars from alternative categories. In fact, as Reed (1972) has noted, almost all models generally predict excellent classification of prototypic patterns, and special measures must be taken in setting up an experiment to distinguish alternative models.

Whether performance on new prototype patterns is better or worse than performance on old training patterns will depend on the component similarity parameters according to the context theory. If similarity is low, few between-category confusions will occur, and training pattern performance will be excellent; if similarity is higher, more between-category confusions will occur, and this will hinder performance on training patterns more than prototype patterns because prototype patterns are unlikely to be highly similar to (confusable with) alternative category exemplars.

These interactions associated with the degree of similarity also provide the basis for accounting for differential retention. Suppose, for example, that the similarity parameters associated with the stimuli in Figure 2 were equal and constant at .10 on an immediate test and (consistent with Assumption 6) that the similarity parameters increased to .40 on a delayed test. Predicted performance on Old Pattern 5 (0011) would drop from 91% correct to 63% correct; while performance on the new pattern (1000) would only drop from 91% correct (Category B responses) to 69% correct on the delayed test.

In general, performance on old patterns will suffer more over a retention interval than performance on new patterns because of the decreasing likelihood that a training pattern probe will successfully access its own stored representation and the associated category assignment information. By a careful selection of training and transfer stimuli, one might be able to produce *increases* in classification performance over a retention interval for certain patterns. As of yet, we are unable to commit ourselves on the issue of whether changes in similarity parameters derive mainly from changes in the availability of component exemplar information or whether changes reflect the forgetting of a strategy involving selective attention.

It seems that in principle, the context theory can account both for excellent classification of prototypic stimuli and for differential reten-

tion of old and new stimuli without positing a (changing) mixture of two types of information. Rather, both results follow from our assumptions concerning retrieval of stored exemplars by similarity. In addition to this apparent parsimony, the context model suggests some new theoretical contrasts that differentiate the context theory from a large class of classification models including prototype models. Before discussing these new contrasts, we characterize this class of models, which we shall call *independent cue models.*

Independent Cue Models

Independent cue models assume that the information entering into category judgments (i.e., overall similarity, distance, or validity) can be derived from an additive combination of the information (similarity, distance, or validity) from the component cue dimensions (Franks & Bransford, 1971; Hayes-Roth & Hayes-Roth, 1977; Reed, 1972). Prototype, average distance, and versions of cue validity and frequency models fall under this domain. First, a general form of an independent cue model will be presented and then the relationship between this general form and particular models will be examined. While the derivation is not completely general, it will hold for binary-valued dimensions.

Consider again Figure 2. A value 1 on any dimension is more closely or more often associated with Category A than Category B, and a value 0 is more likely to be associated with Category B. In other words, a value 1 provides information that a stimulus will probably fall into Category A, while a value 0 provides information that a stimulus will probably fall into Category B. For all of the models to be considered, the probability of classifying a pattern into Category A is assumed to be an increasing function of the weight of evidence favoring Category A and a decreasing function of the weight of evidence favoring Category B. Owing to perceptual salience, selective attention, or particular learning strategy, the component dimensions of a stimulus may not be equally weighted. With these considerations and Figure 2 in mind, one can write an equation for evidence for Category A (E_A) as

$$E_A = (I_{cA} - I_{cB})W_c + (I_{fA} - I_{fB})W_f + (I_{sA} - I_{sB})W_s + (I_{pA} - I_{pB})W_p, \quad (6)$$

where W_c, W_f, W_s, and W_p are the weights associated with the color, form, size, and position dimensions, and where I_{jA} and I_{jB} for a given dimension j ($j = c, f, s$, or p) are, respectively, the information favoring Category A and the information favoring Category B.

The first term, $(I_{cA} - I_{cB})W_c$, consists of a difference in information value $(I_{cA} - I_{cB})$ and an associated weighting factor (W_c). When a pattern is presented to be classified at some point, the value associated with the color dimension (red [1] or blue [0] in our example) will be sampled. If red (denoted 1) is sampled, then $(I_{cA} - I_{cB})$ would be positive, since red is more closely associated with Category A than Category B. If blue (denoted as value 0) were sampled, then $(I_{cA} - I_{cB})$ would be negative, since blue is more closely associated with Category B than Category A. This information from the color dimension would then be weighted by the parameter W_c and added to the weighted information from the form, size, and position dimensions to arrive at the overall evidence favoring Category A. In the example shown in Figure 2, for each dimension, the associated values are equally informative (i.e., probability of Category A given value 1 is equal to the probability of Category B given value 0), which means that $I_{jA} - I_{jB}$ for value 1 is equal to $I_{jB} - I_{jA}$ for value 0. Because of this symmetry, Equation 1 can be rewritten as

$$E_A = W_c I_{ck} + W_f I_{fk} + W_s I_{sk} + W_p I_{pk}, \quad (7)$$

where W_c, W_f, W_s, and W_p are defined as before, and where $I_{j,k}$ for a given dimension j ($j = c, f, s$, or p) and value k ($k = 0$ or 1) is equal to $+1$ if k is 1 and equal to -1 if k is 0. Since for each dimension, the values 1 and 0 are equally informative and since the weights for the various dimensions are arbitrary parameters, Equation 7 represents the general case where for each dimension, the value is sampled to see in which category it is more closely associated. This information is then weighted according to the importance of the dimension. Equation 7 can be used to derive the evidence for Category A for the three A training stimuli in Figure 2 ($E_{A,1}$, $E_{A,2}$ and $E_{A,3}$):

$$E_{A,1} = W_c + W_f + W_s + W_p, \quad (8)$$

$$E_{A,2} = W_c + W_f + W_s - W_p, \quad (9)$$

and

$$E_{A,3} = -W_c - W_f - W_s + W_p. \quad (10)$$

For Stimulus 1, the values on each of the dimensions are more often associated with Category A, and therefore, each of the terms in Equation 8 are positive. For Stimulus 2, the values associated with color, form, and size are also more consistent with Category A; but the value 0 for position was more often associated with Category B, and therefore, W_p in Equation 9 is negative. To the extent that category judgments are based on the position dimension, Stimulus 2 would tend to be inappropriately classified.

From this example, we can see that Stimulus 1 should be easier to learn and classify than Stimulus 2 or 3 because unlike the latter two stimuli, all of its values are associated with positive weights; however, the relative difficulty of learning and classifying Stimulus 2 and Stimulus 3 will depend on the weights associated with the four dimensions.[1] If all dimensions are weighted equally, then Stimulus 2 should be easier to learn and classify than Stimulus 3; and in this case, $E_{A,3}$ would actually be less than zero, which implies that the total information favors the item being sorted into Category B. If a New Stimulus 7 having notation 0101 were introduced, then,

$$E_{A,7} = -W_c + W_f - W_s + W_p, \quad (11)$$

and no clear prediction concerning its classification would be made unless the weights had been determined. If all dimensions were weighted equally, $E_{A,7}$ would be zero, and one would predict that the stimulus would be as likely to be classed as an A or as a B.

In evaluating independent cue models, we will also take into consideration the possibility that category judgments of training stimuli are based on specific item information. This will be represented by a parameter M_i, and it will be assumed that this parameter is added to the other evidence as detailed in Equation 7. In other words, a fifth term, $+M_i$, will be added to Equation 7 for predictions concerning transfer tests involving old training stimuli.

For patterns in Category B, one can either write analogous equations for E_B and assume items are classified into Category B if E_B is greater than E_A, or one simply can note that when E_A is less than zero, a pattern is more closely associated with Category B. In the next paragraphs, particular independent cue models will be described briefly and related to Equation 7.

Prototype Theory

Reed (1972) has provided the most formal and general treatment of prototype theory. The prototype of a category represents its central tendency, and the main idea is that experience with exemplars of a category leads to the development of the prototype, which then provides the basis for classification judgments.

For prothetic (intensity) dimensions, the central tendency is defined as some function of the mean value along each dimension. For metathetic continua, such as color, the prototype is more appropriately defined as some function of the modal values along each dimension, since subjects presented with red and blue stimuli do not act as if they have been presented with purple stimuli (Hirschfeld, Bart, & Hirschfeld, 1975). For binary-valued dimensions, this function will be assumed to be such that if a value is closer to either the mode or the mean associated with a category, the value will be taken as evidence favoring that category. In our example, the modal prototype for Category A is 1111 and for Category B is 0000. The decision rule for responding to new stimuli is to compare the distance from the stimulus to the prototype for Category A and to the prototype for Category B and assign the stimulus to the category whose prototype is closest to the stimulus. These distances are derived by summing the distances along each component dimension. In a generalized form of the model, one allows the distances from the various dimensions to be weighted differ-

[1] Strictly speaking, Equation 7 and our derivation have only been shown to hold for transfer. However, if training stimuli are randomly presented and the development of whichever form of information a model uses is orderly, then the predictions should also hold for learning. One cannot, of course, prove this rigorously, since none of the classification models has been elaborated to account for the details of acquisition.

entially (Reed, 1972) in order to account for selective attention to distinctive features.

To relate prototype theory to Equations 6 and 7, one need only note that distance from the prototype corresponds to information. If a value of some new stimulus on dimension x is closer to the value of Prototype A than Prototype B on that dimension, then this represents evidence that the new stimulus belongs to Category A. Reed (1972) used a particular algorithm to fix the weights of the component dimensions, and Equation 7 provides the more general case where the weights are free parameters.

Other Independent Cue Models

Average distance, cue validity, and frequency models can also be mapped onto Equations 6 and 7. Average distance models assume that people compute the distance of a probe stimulus from all of the stored patterns and assign the probe to the category having the smallest average distance from the probe. Cue validity models propose that people learn the degree to which values on individual dimensions can be used to predict category membership and that category judgments are produced by summing the validity information from the component dimensions (see Reed, 1972, for a more complete presentation of average distance and cue validity models).

Simple frequency models propose that people store the frequency with which the individual attributes of dimensions are associated with each category. In the frequency model assessed by Franks and Bransford (1971), category judgments of probe stimuli are assumed to be derived from summing the frequency with which each of the attributes of the probe have been associated with each of the categories and selecting the category with the higher frequency sum (for further assessment of frequency models, see Hayes-Roth & Hayes-Roth, 1977). Each of the above models assumes that classification is based on an additive summation of component information, and Equations 6 and 7 represent the general case where these component dimensions may be differentially weighted.

Context Model Versus Independent Cue Models

The context model differs from independent cue models in that it assumes an interactive, specifically multiplicative, combination rule for component dimensions. This means that high similarity to particular patterns should determine classification performance more than overall average similarity of a given pattern to other patterns. For independent cue models, average similarity (or its converse, distance) is the only determinant of performance. Previous research does not bear at all directly on this difference between the theories, since most earlier work has concerned itself with the effects of varying the distance of exemplars from the category prototype. The context model predicts that with distance of an exemplar from the prototype held constant, performance will vary with the number of stored exemplars similar to the exemplar in question (category density). Independent cue models are insensitive to such density effects. The experiments in the next section of this article are aimed specifically at this difference between the theories.

Experiment 1

Design and Theoretical Predictions

The design of the first experiment is shown in Figure 3. For each dimension, two of the three values in Category A are 1 and two of the three values in Category B are 0. Thus, each dimension carries some information, but none provides a perfectly valid cue. Certain combinations of cues, such as size 1 and color 1, do provide valid cues. The structure might correspond to an ill-defined concept, where for every dimension, there exists an exception to the rule in each category.

All models will predict that Stimulus 6 and Stimulus 10 (the modal prototypes) should be easiest to learn, while Stimulus 15 should be difficult. Because Stimulus 10 has a high similarity pattern in its own category, while Stimulus 6 has a highly similar pattern associated with the contrasting category, the context model predicts that Stimulus 10 should be easier to learn than Stimulus 6, assuming that

TRAINING STIMULI

"A" STIMULI						"B" STIMULI					
STIMULUS NUMBER	DIMENSION VALUES				RATING	STIMULUS NUMBER	DIMENSION VALUES				RATING
	F	S	C	P			F	S	C	P	
6	1	1	1	1	4.8	10	0	0	0	0	5.2
7	1	0	1	0	4.6	15	1	0	1	1	4.5
9	0	1	0	1	4.8	16	0	1	0	0	4.9

NEW TRANSFER STIMULI

"A"-PREDICTED						"B"-PREDICTED					
STIMULUS NUMBER	DIMENSION VALUES				RATING	STIMULUS NUMBER	DIMENSION VALUES				RATING
	F	S	C	P			F	S	C	P	
5	0	1	1	1	4.3	3	1	0	0	0	3.5
13	1	1	0	1	4.4	8	0	0	1	0	4.0
4	1	1	1	0	3.6	14	0	0	0	1	3.2

Figure 3. Design of Experiment 1. (*F*, *S*, *C*, and *P* refer to the dimensions of form, size, color, and position, respectively. The rating scores are averages over the initial and delayed tests.)

all dimensions are equally salient. Independent cue models predict no difference.

The most interesting predictions arise, however, concerning the new transfer stimuli. These stimuli are grouped into pairs in Figure 3, where each stimulus in a pair has the same weight (though with respect to different categories) on each dimension according to Equation 7. Therefore, independent cue models predict transfer on Stimulus 3 to equal transfer on Stimulus 5, transfer on 8 to equal 13, and transfer on 14 to equal 4. That is, Stimulus 3 should be called an A as often as Stimulus 5 is called a B, and so on. There are also certain paired implications such as if transfer is better on Stimulus 3 than Stimulus 8, then transfer on 5 should be better than on 8 and transfer on both 3 and 5 better than on 13. Between-pair differences might arise if the dimensions were not equally weighted, and the predictions derive from consistency in such weightings.

The context model makes different predictions. One way to see this is to note that Stimuli 3, 8, and 14 are highly similar to one A stimulus and one B stimulus; while Stimuli 5, 13, and 4 are highly similar to two A stimuli and are not highly similar to any B stimuli. Therefore, Stimulus 5 should be classified as an A more readily than Stimulus 3 is classified as a B; similarly, 13 should be better than 8 and 4 better than 14. These predictions can be established more rigorously by writing out prediction equations analogous to Equation 1. Predictions between pairs will depend on the relative salience of the various dimensions.

Method

Subjects. Thirty-two volunteers were solicited through ads in local newspapers. The subjects, men and women ranging in ages from 17 to 30 years, were paid $2.50 for each of two experimental sessions.

Stimulus cards. The 16 stimuli consisted of geometric forms mounted on plain white 12.7 × 20.3 cm index cards. The forms varied along the four binary-valued dimensions of form, size, color, and position. A form was either an equilateral triangle or a circle, either red or green, had a diameter or height of either

1.25 cm or 2.5 cm, and was centered either on the left or right side of the card. For a given subject, six training and six additional transfer stimuli were used, and across subjects all 16 cards were used.

The basic design has already been presented in Figure 3. All subjects were presented with cards in accordance with that general design, but the particular assignment of individual stimulus cards to the abstract notation varied from subject to subject. For example, 1111 might refer to triangle, large, red, and left for one subject; triangle, small, green, and left for another subject; circle, small, green, and right for another subject; and so on. Overall, each card was assigned to a given stimulus notation exactly twice, once when Stimuli 6, 7, and 9 were associated with Category A and once when they were associated with Category B. In other words, the assignment of stimulus cards to conditions and category labels was exactly counterbalanced.

Procedure. The basic procedure in the first session involved initial learning followed by a 5-minute distractor task, which was then followed by transfer tests. In the second session, which was given 1 week later, a 5-minute distractor task was followed by a second transfer test, and then a final classification test was given with the training stimuli visible.

Training. Initial training consisted of up to 20 runs through the list of 6 training stimuli. Subjects were given the following instructions:

> This is an experiment concerned with how we store information in memory. I'm going to present you with some cards belonging to two sets, A and B. At first you will just have to guess which set each card belongs to. But after you make a choice, I'll tell you whether you are right or wrong, so that eventually you should be able to learn which set each card belongs to. Although this task is very difficult, there are no tricks involved, and a given card will always be in the same category.

For training, subjects were shown the cards one at a time and asked to classify them as either A or B. Each card was displayed until the subject responded, immediate feedback concerning correctness or incorrectness was given, and the card continued to be displayed for about 1 sec after feedback was given. Each of the six cards was presented once on each run through the list. The cards were presented in a random order, and there was no obvious break between runs. Training continued until a subject made no errors on two consecutive runs through the list or until the list had been presented 20 times.

Interpolated activity. After the training period, the subjects were asked either to rate the meaningfulness or the pronounceability of consonant–vowel–consonants (CVCs) on a 7-point scale. This activity lasted 5–10 minutes.

Initial transfer test. Subjects were given the following transfer instructions:

> Now I would like you to give a judgment as to which set each card belongs to, but in addition, I'd like you to indicate how confident you are of your judgment. So after you say "A" or "B," say "one" if you feel like you were guessing, "three" if you're sure you're correct, and "two" for somewhere between guessing and sure. Some of the cards you see may be new. If you see one that's new, say "new," and then give your judgment as to which set it belongs to anyway, plus a confidence rating. This time I won't tell you whether you are right or wrong.

Subjects were then shown the 12 cards, one at a time in a random order. Each card was displayed until the subject's judgment and confidence rating was completed. No feedback was given concerning either the judgment or whether the pattern was old or new.

One week later, subjects returned to the laboratory for additional tests. First, subjects rated either the meaningfulness or pronounceability of CVCs for 5–10 minutes. Subjects who had originally rated meaningfulness were now asked to rate pronounceability and vice versa. Then, subjects were given a transfer test involving the same 12 stimulus cards and the identical instructions they had experienced a week earlier. The only difference was that a new random order was used for the stimulus presentation. Again no feedback was given.

Final classification. Immediately following the transfer task, the experimenter laid out the three A training cards and the three B training cards in two groups in front of the subjects. They were told that these were the training cards and then were shown the six new cards, one at a time, and asked with which group they thought the card went. In addition, subjects were asked to give a confidence rating on their judgment. No feedback was given on judgments, since feedback would have been inappropriate because there were no right or wrong answers.

Results

Learning. All but 5 of the 32 subjects met the learning criterion. Mean errors on Stimulus Numbers 6, 7, 9, 10, 15, and 16 were 3.6, 4.7, 4.4, 3.1, 4.9, and 3.8, respectively. As expected by all theories, the modal prototypes—Stimulus 10 and Stimulus 6—were easiest to learn; while Stimulus 15, which according to independent cue theories could only be mastered by the use of specific item information, proved to be the most difficult to master.

Transfer. The recognition data reveal a modest ability to discriminate training and new transfer stimuli. On the first transfer test, the hit rate was .98 and the false alarm rate (saying "old" to a new stimulus) was .69; on the second test given a week later, the hit rate was .96 and the false alarm rate was .75. An analysis of variance on the new stimuli indicated that the effect of time did not reach statistical significance nor were there significant differences in recognition between new transfer stimuli.

The transfer category responses and confidence ratings were transformed into a 6-point rating score as follows: 1 = high confidence error, 2 = medium confidence error, 3 = guessing error, 4 = guessing correct, 5 = medium confidence correct, and 6 = high confidence correct. For new stimuli, where correctness and incorrectness are inappropriate, the scale is arbitrarily defined with respect to what would be predicted from an independent cue model with each dimension equally weighted. This definition simply provides a consistent standard and does not favor one set of theories over the other.

The transfer rating data are shown in Figure 3, a mean rating of 3.5 representing chance or nondifferential classification. The scores in the figure are an average of the ratings on the initial and delayed transfer tests. The greatest interest lies in performance on the new transfer stimuli. On all three tests, Stimulus 5 was rated higher than Stimulus 3, 13 was rated higher than 8, and 4 was rated higher than 14. These differences are predicted by the context model and are contrary to the predictions of the general independent cue model. There are also large between-pair differences in rating scores, suggesting that all dimensions were not equally salient.

An analysis of variance on rating scores conducted for the new transfer stimuli showed that both the effect of sets (4, 5, and 13 vs. 3, 8, and 14), $F(1, 30) = 4.36$, $MS_e = 6.21$, $p < .05$, and the effect of pairs (3 and 5 vs. 8 and 13 vs. 4 and 14), $F(2, 60) = 5.76$, $MS_e = 3.57$, $p < .01$, were significant. Neither the effects of tests nor the interactions involving tests were reliable, although both the Tests $\times$ Sets, $F(1, 30) = 3.29$, $MS_e = 1.40$, $p < .10$, and the Tests $\times$ Pairs, $F(2, 60) = 2.58$, $MS_e = 2.58$, $p < .10$, interactions approached significance. On the final classification test, both the effect of sets, $F(1, 30) = 6.79$, $MS_e = 2.86$, $p < .01$, and the effect of pairs, $F(2, 60) = 3.18$, $MS_e = 4.16$, $p < .05$, were significant.

Ratings for the training stimuli were quite high on the initial test and showed a sharp drop (from an average of 5.3 to an average of 4.4) over the 1-week retention interval. The new transfer stimuli did not show this sharp drop between Test 1 and Test 2, changing from an average of 3.9 to an average of 3.8. An analysis of variance was conducted using old versus new stimuli and tests as factors to assess differential retention. The effects of old versus new stimuli, $F(1, 30) = 32.5$, $MS_e = 31.0$, $p < .01$, tests, $F(1, 30) = 23.03$, $MS_e = 16.0$, $p < .01$, and the interaction of these two factors, $F(1, 30) = 24.3$, $MS_e = 10.0$, $p < .01$, were significant. The interaction suggests that old patterns were forgotten more rapidly than new patterns, a result that has been obtained frequently in this area. One must be careful in inferring differential retention based on a statistically significant interaction, particularly since ratings on several of the transfer stimuli are near chance. If we consider only transfer stimuli showing clear above-chance ratings (5, 8, and 13) versus the old stimuli and subtract 3.5 from each score and divide by the base value (first transfer test) to get a measure of percentage retained, old patterns show a mean of 41% retention, while the three new stimuli under consideration show a mean of 64% retention. In addition, if we disregard the rating scores and consider only percentage correct classifications, Stimulus Numbers 5, 8, and 13 again show a better retention than training stimuli over the 1-week interval (62% vs. 39% retained). Therefore, there is at least modest evidence for differential retention of old and new stimuli.[2]

[2] Two surprising results were that Stimulus 15, which should have been difficult to classify, was rated higher than Stimulus 16 on the initial test, and that Stimulus 4 and Stimulus 14 received such low ratings. We suspect that this result was produced by a specialized strategy involving the position dimension. Specifically, it appears that subjects normally pay little attention to position, but if two stimuli differing only in position fall into distinct categories, subjects may use the strategy of assigning all stimuli differing only in position to distinct categories. In the design of Experiment 1, there are two stimulus pairs differing only in position (Stimuli 7 and 15 and Stimuli 9 and 16), and they both were assigned to different categories. If such a strategy carried over into transfer, then performance on Stimulus 4 and Stimulus 14 would suffer, because Stimulus 14 differs from Stimulus 10 only in position, while 4 differs from 6 only in position; this strategy would lead to 14 being classified as an A and 4 being classified as a B. Although the use of this strategy undermines any attempts to fit the transfer data quantitatively, the qualitative differences between Stimuli 3 and 5 and between 8 and 13 would not be affected, since none of these four stimuli differs from any training stimulus

Discussion

The transfer results of Experiment 1 were in accord with predictions derived from the context model but inconsistent with independent cue models. This finding held for an immediate transfer test, a similar transfer test given 1 week later, and on a final classification task where the training cards were displayed and subjects were asked to sort the new cards into categories based on the visible training stimuli.

Aside from questions about generality, two serious reservations might be posed concerning Experiment 1. The most salient objection is that according to the structural constraints imposed in some experiments with artificial categories (e.g., Reed, 1972; Reed & Friedman, 1973), the A and B stimuli did not comprise proper categories because Stimulus 15 (1011) belongs in the category where the value 1 appears most often (Category A). More formally, one would argue that the categories were not separable by a linear discriminant function in the sense that no additive combination of dimension weights and values could be used to correctly classify the training stimuli (see Reed, 1972; Sebestyen, 1962). Although the training stimuli could be unambiguously sorted by an independent cue model if we used the parameter M_i for specific item information, this objection must be taken seriously because independent cue models may operate only when linear separability holds. On the other hand, there is no evidence to indicate that natural categories conform to linear separability.

A second objection is that Experiment 1 is biased in the sense that support for independent cue models would have amounted to embracing the null hypothesis, which generally speaking, is not an attractive strategy. One would prefer a situation where the context theory and independent cue models make different predictions, neither of which amount to predicting the absence of differences.

Experiment 2 is directed toward answering both objections. The categories employed conform to linear separability, and the theories make distinctly different qualitative predictions.

only in position. Because of the potential use of this strategy, position was not used as a stimulus dimension in subsequent experiments.

Experiment 2

Design and Theoretical Predictions

The structure of Experiment 2 is shown in Figure 4. As in Experiment 1, the value 1 is likely to be associated with Category A and the value 0 with Category B for each of the dimensions. Unlike the other experiments, however, a linear discriminant function may be used to separate the two categories. One easy way to see this is to note that if the form dimensions were given zero weight, then at least two of the other three dimension values for each stimulus are appropriate for the category. The categories are not well defined, since no specific feature values are even necessary for category membership.

The main prediction of interest concerns Stimuli 4 and 7. Since the modal prototype is 1111, Stimulus 4 must be at least as close as 7 is to the prototype, no matter how the dimensions are weighted. More generally, all independent cue models will predict that Stimulus 4 will be easier to learn than 7 because for the only dimension where the two stimuli differ, 4 will have a positive weight and 7 a negative weight (unless $W_f = 0$). Unlike Experiment 1, an overall bias favoring one category response over the other will not change this prediction.[3]

In contrast, the context model predicts that Stimulus 7 should be easier to learn than Stimulus 4 because the effect of number of highly similar patterns is the most important factor in performance. Stimulus 7 is highly similar (i.e., differs in only one dimension) to two other Category A patterns (Stimuli 4 and 15) but is not highly similar to any Category B patterns. Stimulus 4, on the other hand, is highly similar to one Category A pattern (Stimulus 7) and to two Category B patterns (Stimuli 2 and 12) and hence should be more difficult to learn. This prediction is not completely parameter free, but it holds over a large range of parameter values, and param-

[3] The symmetry assumption embodied in Equation 7 does not hold for the dimensions in the design of Experiment 2. For example, for color, the probability of Category A given a value 1 is .80, while the probability of Category B given a value 0 is .75. The asymmetries are extremely small, however, and the qualitative conclusions drawn will not hinge on the symmetry assumption.

TRAINING STIMULI

"A" STIMULI

Stimulus Number	Dimension Values C	F	S	N	Fe	Rating
4	1	1	1	0	4.9	4.8
7	1	0	1	0	3.3	5.4
15	1	0	1	1	3.2	5.1
13	1	1	0	1	4.8	5.2
5	0	1	1	1	4.5	5.2

"B" STIMULI

Stimulus Number	Dimension Values C	F	S	N	Fe	Rating
12	1	1	0	0	5.5	5.0
2	0	1	1	0	5.2	5.1
14	0	0	0	1	3.9	5.2
10	0	0	0	0	3.1	5.5

NEW TRANSFER STIMULI

Stimulus Number	Dimension Values C	F	S	N	Rating A-Predicted	Rating B-Predicted
1	1	0	0	1	3.7	
3	1	0	0	0		4.4
6	1	1	1	1	5.3	
8	0	0	1	0		4.1
9	0	1	0	1	3.3	
11	0	0	1	1	4.1	
16	0	1	0	0		4.9

Figure 4. Design of Experiment 2. (*C*, *F*, *S*, and *N* refer to the dimensions of color, form, size, and number, respectively. The stimulus numbers are carried over from Experiment 1. *Fe* refers to average errors during learning. Rating scores may vary from 1 to 6, with 3.5 representing chance [nondifferential] classification.)

eter values that would alter this prediction would place numerous other testable constraints on the data.

Method

Subjects. Thirty-two volunteers were solicited through ads in local newspapers. The subjects, men and women ranging in ages from 17 to 30 years, were paid $2.50 for the experimental session. The subjects had not participated in the first experiment.

Stimuli. Sixteen stimulus cards with geometric forms drawn on them were used. Nine cards were used in training and seven additional cards were used in transfer. The geometric forms were like those from the preceding experiment, except that the dimension of number was substituted for the dimension of position. The number dimension was represented by either a single geometric form centered on the card or by two geometric forms each centered on their respective halves of the card.

The assignment of abstract notation to individual stimulus cards varied from subject to subject exactly as in the first experiment. That is, the assignment of stimulus cards to conditions and category labels was exactly counterbalanced.

Procedure. The procedure followed that used in the first part of Experiment 1: initial training, followed by a 5–10-minute interpolated activity, followed by a transfer task involving both training and new transfer stimuli.

The instructions for training were those used in Experiment 1. Training consisted of up to 16 runs through the list of 9 training stimuli with a learning criterion of 1 errorless run. Other procedural details followed those of Experiment 1, including the interpolated activity and the transfer test instructions and procedure.

Results

Learning. The learning task was of moderate difficulty; 19 of the 32 subjects learned the classification task within the maximum limit of 16 runs. Overall, subjects averaged 18% errors on the last run through the list, but virtually all subjects showed some improve-

ment with practice. Mean errors for each stimulus are shown in Figure 4. Contrary to the independent cue models and consistent with the context model, Stimulus 4 proved to be more difficult than Stimulus 7.

An analysis of variance on errors showed that the effect of stimuli, $F(8, 240) = 4.29$, $MS_e = 6.27$, $p < .01$, was statistically significant. A Duncan's multiple-range test using the .05 significance level indicated that Stimuli 2, 4, 12, and 13 were associated with more errors than 7, 10, and 15; Stimulus 5 had more errors than 10; and Stimulus 12 had more errors than 14. A planned direct t test of Stimulus 4 versus Stimulus 7 was, of course, also significant ($t_{31} = 3.71$, $p < .01$, two-tailed).

Transfer. The overall hit rate for old stimuli was .99, while the false alarm rate for new stimuli was .87. Stimulus 6, the modal prototype for Category A, was never correctly rejected as new. Although no other new stimulus had a zero correct rejection rate, an overall analysis of correct rejections indicated that the effect of stimuli was not reliable, $F(6, 186) = 1.51$, $MS_e = .10$, $p > .10$.

The transfer rating data appear in Figure 4. Stimulus 7 received a higher mean rating than Stimulus 4, but the effect fell short of statistical significance ($t \cong 1$). Old stimuli generally received higher ratings than new stimuli, and the transfer rating data were in reasonable accord with both theories.

Because so many subjects failed to meet the learning criterion, the data for learners and nonlearners were considered separately. No systematic differences were noted. Both learners and nonlearners made more errors on Stimulus 4 than Stimulus 7, and there was a similar overall pattern of performance in the two groups. Learners averaged 88% correct on the transfer test, nonlearners averaged 78% correct, and the rank-order correlation of the classification probabilities for the two groups was high and positive (+.62).

Detailed fits of theories and data. Although the focus has been on qualitative comparisons of the models, it is also of interest to see how well they fit the data overall. For the context model prediction, equations analogous to Equations 1 and 4 were written for each stimulus pattern in terms of the similarity parameters associated with the four dimensions. To estimate the similarity parameters, programs were written to minimize the average absolute deviation of predicted and observed classification probabilities. No distinction was made between old training patterns and new patterns in applying the model.

For the error data predictions, equations analogous to Equation 1 were again used, and programs were written to search the parameter space to maximize the correlation between ranked evidence values (E_A or E_B) and the ranked error data. To make appropriate comparisons with the independent cue model, the correlation between ranked evidence values and ranked classification data was also calculated. The ranked evidence values were derived from the parameter values used to generate predictions of classification probabilities.

A similar procedure was used for the class of independent cue models. Equation 7 was specialized for each pattern, and the parameter space was searched to find values that would maximize the correlation between the ranked evidence values (E_A and E_B) and ranked errors and separately between the ranked evidence values and the ranked transfer classification probabilities. For the error data, parameters corresponding to the weight of each dimension were used; and for the transfer data, a fifth parameter for the weight of specific item information was also estimated. Since the dimension weighting parameters involve relative weightings, there really are only three independent dimension weight parameters.

The correlation between predicted and observed ranked errors was +.99 for the context model and +.81 for the independent cue model. The parameters associated with these correlations were $c = .14$, $f = .18$, $s = .16$, and $n = .14$ for the context model and $W_c = .40$, $W_f = .10$, $W_s = .50$, and $W_n = .20$ for the independent cue model. Table A1 in the Appendix shows predicted and observed error ranks.

Observed classification probabilities and predictions based on the context model are shown in Table 1. With the exception of Stimulus 15, most of the predictions are quite close: The rank-order correlation between predicted and observed classification probabilities was +.81. Note that data from both old and new stimuli are all well predicted, even though no special

Table 1
Predicted and Observed Classification Probabilities for the Transfer Task of Experiment 2

Stimulus number	Classification probability: Observed	Classification probability: Predicted
Training stimuli		
4A	.78	.79
7A	.88	.94
15A	.81	.97
13A	.88	.86
5A	.81	.86
12B	.84	.76
2B	.84	.76
14B	.88	.93
10B	.97	.97
New transfer stimuli		
1A	.59	.64
6A	.94	.93
9A	.50	.57
11A	.62	.64
3B	.69	.61
8B	.66	.61
16B	.84	.87

Note. The letters A and B define the category with respect to which classification proportions are scored. Each observed proportion is based on 32 observations.

assumptions are made to distinguish new and old stimuli. For the independent cue model, the rank-order correlation was +.79, indicating that the overall fit was approximately as good as that of the context model. The parameters associated with the classification data were $c = .16$, $f = .16$, $s = .18$, and $n = .14$ for the context model and $W_c = .38$, $W_f = .10$, $W_s = .40$, $W_n = .20$, and $M_i = .35$ for the independent cue model. Table A2 in the Appendix summarizes these rank-order predictions.

Discussion

The results on errors, but not on classification, favor the context model. A stimulus close to the modal prototype proved to be more difficult to learn and classify than one farther away. In addition, quantitative predictions of the context model were in good agreement with the data, and overall ranked-order correlations between predicted and observed error and classification data were higher for the context model than for the independent cue model. Apparently, difficulties for the independent cue model are not confined to situations where the classes are not linearly separable.

So far, we know little about the generality of these results to other stimulus domains. The next experiment uses the same design as employed in Experiment 2, but now the stimuli are Brunswik faces rather than geometric forms. These face stimuli were selected because the clearest quantitative evidence favoring prototype models, a subset of independent cue theories, was derived from studies using these stimuli (Reed, 1972).

Experiment 3

The structure of Experiment 3 mirrors that of Experiment 2 as shown in Figure 4. The two categories are linearly separable, and the qualitative focus again is on the learning and classification of Stimulus 4 versus Stimulus 7, since distance from the category central tendency and the number of highly similar training patterns are placed in conflict. In addition to learning and classification data, Experiment 2 included an additional training phase where response latency was the dependent variable of interest. We will assume for both theories that the greater the evidence favoring a response to a training stimulus, the faster will be the response. The major difference between Experiments 2 and 3 is that Experiment 3 employs Brunswik faces rather than geometric forms as stimuli.

Method

Subjects. Thirty-two volunteers were solicited through ads in local newspapers. The subjects, men and women ranging in ages from 17 to 30 years, were paid $2.50 for the experimental session, which lasted approximately 50 minutes. No subject had participated in either of the first two experiments.

Stimuli. The stimuli were Brunswick faces displayed on an approximately 27 × 34 cm visual display screen (Digital Equipment Corp. VR-17 cathode-ray tube screen) linked to a PDP-11 computer. The face outlines were 13.5 × 11.5 cm and centered on the screen. The faces differed in nose length, mouth height, eye separation, and eye height, which were the four dimensions that had been varied previously by Reed (1972). The nose was either a 1.5-cm or a 3.0-cm vertical line centered within the face outline. The mouth was a

4.0-cm horizontal line, which was either 1.5 cm or 3.0 cm from the chin line. The eyes were 1 cm × 2.5 cm and were separated either by 1.5 cm or 3.5 cm (measuring from inner edges). Finally, the eyes were either 2.5 cm or 5 cm from the top of the face outline (measured to the top edge of the eye). The two possible values on each of the four dimensions were combined to produce 16 distinct stimuli.

Categories were constructed in accordance with the design shown in Figure 4. Moving from left to right, the dimensions of eye height, eye separation, nose length, and mouth height were substituted for the dimensions of color, form, size, and number.

Procedure. The experiment had three main phases: initial training, transfer test, and speeded classification. Initial training consisted of up to 32 runs through the list of 9 training stimuli with a learning criterion of 1 errorless run. The basic trial sequence was as follows: A face appeared on the screen and remained on until the subject pressed either the button marked "A" or the button marked "B," which occupied the lower left and lower right corners, respectively, of a 4 × 4 button response box; the face remained on the screen for 2 sec, while feedback ("Correct"; "No, that is A"; or "No, that is B") was displayed below the face; a 1-sec interstimulus interval ensued. The initial training instructions paralleled those of the first two experiments.

Transfer tests immediately followed initial training and mirrored the procedures used in the earlier experiments. The only difference was that two randomly scrambled runs through the 16 possible faces were given, staying on the screen until the confidence judgment was completed. The interstimulus interval was as before, but no feedback concerning either recognition or classification was given.

After these transfer tests, subjects were given an additional 16 runs through the nine training stimuli. Presentation and feedback were exactly as in initial training. The only difference was that subjects were told we were now recording response latencies and that we wanted them to respond as fast as they could without making errors.

Results

Learning. Apparently, faces were harder to discriminate than geometric stimuli, since only 14 of the 32 subjects met the learning criterion within the maximum of 32 runs. Mean errors for each stimulus are shown on the left side of Table 2. Stimulus 7 again proved to be easier to learn than Stimulus 4, contrary to the qualitative predictions of the independent cue model. This difference is not large, however, and though the overall effect of stimuli is significant, $F(8, 240) = 31.5$, $MS_e = 17.2$, $p < .001$, the difference between errors on Stimulus 7 and Stimulus 4 was not significant. On the other hand, if we consider only the data of the 14 subjects who mastered the task, the difference is reliable ($t_{13} = 2.84$, $p < .02$, two-tailed).

Table 2
Mean Errors and Mean Rating Scores for Experiment 3

Learning		Transfer	
Stimulus number	Mean errors	Stimulus number	Mean rating
4	5.5	4A	5.2
7	4.2	7A	5.2
15	2.8	15A	5.1
13	11.9	13A	4.7
5	8.2	5A	4.1
12	15.2	12B	4.3
2	12.9	2B	4.1
14	6.6	14B	5.1
10	4.4	10B	5.2
		1A	4.0
		3B	3.7
		6A	5.2
		8B	4.7
		9A	2.5
		11A	2.7
		16B	4.8

Note. The mean rating score is with respect to the category label (A or B) associated with each stimulus in the table. Rating scores may vary from 1 to 6, with 3.5 representing chance (nondifferential) classification.

Transfer. Recognition scores were low: The probability of saying "old" given an old face was .82, while the probability of saying "old" given a new face was .79. Neither the old–new difference nor individual stimulus differences were significant.

The transfer rating data appear on the right side of Table 2. The mean rating scores for Stimulus 4 and Stimulus 7 were nearly identical. Generally, old stimuli received higher ratings than new stimuli; but New Stimulus 6, the modal prototype for Category A, had as high a rating as any of the training stimuli. Detailed fits of the models to these data will be considered shortly.

Speeded classification. Since fewer than half the subjects met the learning criterion during initial training, the speeded classification data were associated with a relatively high proportion of errors (17% overall). The mean error and mean correct latency data are shown in Table 3. As can readily be seen, the correlation

Table 3
Mean Errors and Mean Correct Latency for Each of the Training Stimuli in the Speeded Classification Phase of Experiment 3

Stimulus number	Mean latency	Mean errors
4	1.11	2.12
5	1.34	5.19
7	1.08	1.00
13	1.27	3.12
15	1.07	.18
2	1.30	4.50
10	1.08	.88
12	1.37	5.75
14	1.13	1.38

between the error and latency data is very high. Apparently, differences in error scores do not arise from a simple trade-off of speed and accuracy. Stimulus 7 was associated with fewer errors and faster correct latencies than Stimulus 4. Again these differences are small and unreliable. Each of the 14 subjects who met the learning criterion, however, responded faster to Stimulus 7 than to Stimulus 4. This latency difference is significant ($t_{13} = 3.67$, $p <$ .01, two-tailed) and corresponds to a small difference in error rates, Stimulus 4 being associated with 2.2% errors and Stimulus 7 with 1.3% errors.

Although learners showed a larger difference between Stimulus 4 and Stimulus 7 for both errors and speeded classification, the classification performance of the two groups was highly comparable. Learners averaged 95% correct on transfer tests, nonlearners averaged 78% correct, and the rank-order correlation between the classification probabilities for the two groups was high and positive (+.88).

Detailed fits of theories and data. The overall fit of the context model and the independent cue models was assessed in the same manner as in Experiment 2. For the context model, equations analogous to Equations 1 and 4 were used; and for the general independent cue model, functions based on Equation 7 were employed.

For errors during learning, the correlations between predicted and observed ranked errors were +.95 for the context model and +.88 for the independent cue model. The parameters associated with these predictions were eye height (*eh*) = .20, eye separation (*es*) = .20, nose length (*nl*) = .15, and mouth height (*mh*) = .35 for the context model and $W_{eh} = .40$, $W_{es} = .00$, $W_{nl} = .50$, and $W_{mh} = .20$ for the independent cue models. Table A3 in the Appendix shows the predicted and observed error ranks for each model.

For transfer, observed classification probabilities and predicted values based on the context model are shown in Table 4. The average deviation of predicted and observed values is less than 4%, with the largest deviation being 9 percentage points. The rank-order correlation between predicted and observed classification probabilities was +.92. For the independent cue model, the best rank-order correlation was +.89, nearly as high as that of the context model. The parameters associated with the classification data were *eh* = .00, *es* = .20, *nl* = .10, and *mh* = .40 for the context model and $W_{eh} = .40$, $W_{es} = .10$, $W_{nl} = .35$, $W_{mh} = .14$, and M_i (weight of specific item information) = .12 for the independent

Table 4
Predicted and Observed Classification Probabilities for the Transfer Task in Experiment 3

Stimulus number	Classification probability: Observed	Classification probability: Predicted
Training stimuli		
4A	.97	.93
7A	.97	.99
15A	.92	.99
13A	.81	.73
5A	.72	.70
12B	.67	.65
2B	.72	.72
14B	.97	.99
10B	.95	.99
New transfer stimuli		
1A	.72	.81
6A	.98	.95
9A	.27	.24
11A	.39	.48
3B	.44	.45
8B	.77	.81
16A	.91	.90

Note. The letters A and B define the category with respect to which the classification proportions are scored. Each observed proportion is based on 64 observations.

cue model. Table A4 in the Appendix summarizes these rank-order predictions.

For speeded classification, best-fitting parameters for the context model produced rank-order correlations of +.98 with the error data and +.95 with the latency data. Respective correlations for the independent cue model were +.92 and +91. For both models, the same parameters produced best fits to both the error data and the latency data, the values being $eh = .00$, $es = .22$, $nl = .10$, and $mh = .40$ for the context model and $W_{eh} = .40$, $W_{es} = .00$, $W_{nl} = .50$, and $W_{mh} = .20$ for the independent cue model. Table A5 in the Appendix shows the predicted and observed rankings for these two models.

Discussion

The results again favored the context model over the independent cue model. Learners made more errors on and classified more slowly a stimulus close to the prototype (Stimulus 4) than one farther away (Stimulus 7). Quantitative predictions of both models were in good accord with the data, but in each case, the correlation between predicted and observed ranks was higher for the context model than for the independent cue model. Apparently, the advantage of the context model over the independent cue model is not limited to studies using geometric stimuli or metathetic stimulus dimensions.

Experiment 4

The retrieval process implied by the context model is straightforward. Although Equation 1 seems to require the subject to do considerable computation to develop the various similarity estimates, actually this need not be so. One could assume that judgments are based on the first pattern retrieved, with the similarity parameters determining the likelihood that particular stimuli will be the first retrieved. An obvious advantage of assuming that performance is based on the initial pattern or patterns retrieved is that one would like to generalize the context model to classification tasks where a large number of exemplars may be involved, and the plausibility of an exhaustive retrieval plan would become quite strained.

If classification judgments are based on the first pattern retrieved, on what are confidence ratings based? A strong possibility is that confidence ratings are based on how quickly or easily the first pattern is retrieved. This would imply that confidence ratings for new transfer stimuli should increase with the number of highly similar patterns, regardless of whether the same classification response is associated with the various patterns. Our data give some support to this idea. For example, judgments of Stimulus 16 in Experiment 2 received a high confidence rating 16 times, and in each case, the category assignment was a B; judgments on Stimulus 9 received a high confidence rating 17 times, but the stimulus was assigned to Category A only 6 of those 17 times. Unfortunately, the number of patterns highly similar to a new transfer stimulus scarcely varied within experiments. Stimulus 6 in Experiment 2 had more similar patterns (four vs. three) than the other transfer stimuli and was most likely to receive a high confidence rating and least likely to receive a low confidence rating. When a stimulus was judged to be new, confidence ratings were low—only 16% of the time were subjects highly confident, and half of the time they reported that they were guessing.[4]

New–old recognition may also be based on how quickly or easily the probe retrieves information and, therefore, may be based on the number of highly similar training patterns. In Experiment 2, a transfer stimulus (Number 6) had more highly similar training patterns than the other stimuli. A direct test indicates that this stimulus was more likely to be judged as old than the other stimuli ($t_{31} = 4.38$, $p < .01$). Unfortunately, the first three experiments provide little variation in the number of similar training patterns. Experiment 4 was

[4] Although confidence judgments depend on the number of highly similar exemplar patterns and hence similarity parameters, similarity cannot be the sole determinant of confidence judgments. The context model represents forgetting in terms of increases in similarity, and it is extremely unlikely that forgetting is accompanied by corresponding increases in confidence rating. Presumably, additional factors such as changes in context determine the general accessibility of stored exemplars, and this general accessibility decreases over time.

TRAINING STIMULI

"A" STIMULI

Stimulus Number	Dimension Values C	F	S	N	Fe	Rating
2	0	1	1	0	5.1	4.5
4	1	1	1	0	3.8	4.7
5	0	1	1	1	4.1	5.0
7	1	0	1	0	4.3	5.1
13	1	1	0	1	4.9	4.8
15	1	0	1	1	4.3	5.1

"B" STIMULI

Stimulus Number	Dimension Values C	F	S	N	Fe	Rating
1	1	0	0	1	5.0	4.8
3	1	0	0	0	5.6	4.3
8	0	0	1	0	5.8	4.3
11	0	0	1	1	4.9	4.5
16	0	1	0	0	4.8	4.4

NEW TRANSFER STIMULI

Stimulus Number	Dimension Values C	F	S	N	Rating A-Predicted	Rating B-Predicted
6	1	1	1	1	4.6	
9	0	1	0	1	3.8	
10	0	0	0	0		4.4
12	1	1	0	0	3.6	
14	0	0	0	1		3.9

Figure 5. Design of Experiment 4. (*C*, *F*, *S*, and *N* refer to the dimensions of color, form, size, and number, respectively. The stimulus numbers are carried over from the first two experiments. *Fe* refers to average errors during learning. Rating scores may vary from 1 to 6, with 3.5 representing chance [nondifferential] classification.)

designed to see whether recognition varies with the number of similar training patterns, as would be expected if recognition were based on the ease with which stored information could be accessed. The experiment also provides still another test of the two main classification models.

The design of Experiment 4 is shown in Figure 5. As before, the value 1 tends to be associated with Category A and the value 0 with Category B. The two categories are separable by a linear discriminant function, as can readily be seen by assuming that the number dimension receives no weight at all, for in that case, two of the three values of each training stimulus match the modal value for their category. The main focus of Experiment 4 concerns new–old recognition, specifically on the predictions of the context model that the more stored exemplars similar to the probe, the less likely the probe will be called new. This should hold for both old and new probe stimuli. For example, Stimulus 16 is highly similar to only one other training pattern (2), while Stimulus 2 is highly similar to four training stimuli (4, 5, 8, and 16); therefore, Stimulus 16 should be more likely to be called new than Stimulus 2. Likewise, New Stimulus 14 is highly similar to two B training stimuli (1 and 11), while Stimulus 12 is highly similar to two A and two B training patterns, so subjects should call Stimulus 14 new more often than Stimulus 12.

No general predictions concerning recognition for independent cue models are available. For new patterns, one might speculate that the farther a probe is from a prototype pattern, the more likely it would be called new. In this case, Stimulus 6 and Stimulus 10, the modal prototypes, should rarely be called new; if the four dimensions are equally weighted, Stimulus

12 should be called new more often than Stimulus 14.

Although the focus of Experiment 4 was on recognition, one should note that the two theories will differ also in their predictions concerning classification. According to the independent cue models, each member of the following stimulus pairs should produce equivalent performance: 2 and 1, 5 and 3, 13 and 8, 15 and 16, and 6 and 10. The context model generally predicts that the first member of each of these pairs will be classified more accurately, partly on the basis of high-similarity exemplars and partly on the basis of the extra Category A pattern that provides an additional opportunity for an A pattern to be retrieved by a probe.

Method

Subjects. Thirty-two volunteers were solicited through ads in local newspapers. The subjects, men and women ranging in ages from 17 to 30 years, were paid $2.50 for the experimental session. The subjects had not participated in any of the first three experiments.

Stimuli. Sixteen stimulus cards with geometric forms drawn on them were used. Eleven cards were used in training, and five additional cards were used during transfer tests. The same stimulus dimensions were used as in Experiment 2, but new values for the color and form dimensions were employed. The forms were either equilateral triangles or hexagons, and they were colored either yellow or purple. The values for the size and number dimensions were as before.

The assignment of abstract notation to individual stimulus cards varied from subject to subject exactly as in the preceding experiments. The assignment of stimulus cards to stimulus conditions and category labels was exactly counterbalanced.

Procedure. Original training was followed by a 5–10-minute interpolated activity, then a transfer test involving both training and new transfer stimuli. Training instructions and the interpolated activity were as in the first two experiments. Training consisted of up to 16 runs through the list of 11 training stimuli with a learning criterion of 1 errorless run. Other procedural details followed those of Experiment 2 including the interpolated activity and the transfer test instructions. The only difference was that the 16 stimulus cards were presented twice during transfer in 2 randomized runs.

Results

Learning. Half of the subjects met the learning criterion within the 16-run limit, and only two subjects failed to show improvement with practice. Mean errors for each stimulus are shown in Figure 5. There was not much variation in errors, and an analysis of variance on errors did not reach statistical significance ($F < 1$).

Transfer. For recognition, overall, the probability of saying "old" given an old pattern was .95, and the probability of saying "old" given a new pattern was .72. On both old and new patterns, the probability of saying "old" increased with the number of training patterns highly similar to the probe. For purposes of analysis, patterns were classified as having large, medium, or small numbers of highly similar training patterns, with large, medium, and small corresponding to four, three, and two highly similar patterns for a new stimulus and corresponding to four, three or two, and one highly similar patterns for an old probe stimulus, respectively. The probabilities of new responses were .03, .05, and .10, respectively, for large, medium, and small numbers of similar patterns for old stimuli and .24, .28, and .38 for corresponding new patterns. These results are consistent with the predictions of the context model. An analysis of variance indicated that both the effects of old versus new probes, $F(1, 31) = 29.8$, $MS_e = 5.08$, $p < .01$, and the effects of number of highly similar training patterns, $F(2, 62) = 4.65$, $MS_e = 2.22$, $p < .05$, were significant.

The rating scores from classification tests are shown in Figure 5. Generally speaking, old stimuli received higher ratings than new stimuli, and of the new stimuli, Modal Prototypes 6 and 10 received higher ratings than other new stimuli. An analysis of variance indicated that the effect of stimuli was significant, $F(15, 450) = 3.08$, $MS_e = 6.29$, $p < .01$. The relation of the classification results to the theories will be brought out when the detailed fits of theory and data are considered.

In view of the large number of nonlearners, the data for learners and nonlearners were considered separately. The recognition data revealed that learners were more accurate in detecting new patterns (probability of saying "new" given new was .35 for learners and .22 for nonlearners) and were slightly more likely to call an old pattern new (.07 vs. .04); also, the effects of the number of highly similar patterns were more clear for learners than nonlearners. On the classification tests, learners

Table 5
Predicted and Observed Classification Probabilities for the Transfer Task of Experiment 4

Stimulus number	Classification probability: Observed	Classification probability: Predicted
Training stimuli		
2A	.80	.74
4A	.78	.90
5A	.86	.80
7A	.83	.72
13A	.72	.78
15A	.80	.74
1B	.70	.70
3B	.64	.73
8B	.73	.73
11B	.78	.72
16B	.69	.69
New transfer stimuli		
6A	.89	.88
9A	.56	.56
10B	.78	.83
12A	.62	.62
14B	.64	.59

Note. The letters A and B define the category with respect to which the classification proportions are scored. Each observed proportion is based on 64 observations.

averaged 84% correct, nonlearners averaged 68% correct, and the correlation between the classification performance in the two groups was modest and positive (+.41). No systematic differences between the two groups were noted, and the low correlation apparently derives from the low overall performance of nonlearners.

Detailed fit of theories and data. Since the effect of stimuli on errors during learning was not significant, no attempt was made to fit the acquisition data. Fits to the classification data were attempted by using equations analogous to Equations 1 and 4 for the context model and Equation 7 for the independent cue model.

Observed classification probabilities and predicted values based on the context model are shown in Table 5. Overall, the average deviation of predicted from observed values is less than 5%, with the largest deviations being the overprediction of performance on Stimulus 4 and the underprediction of performance on Stimulus 7. The rank-order correlation of predicted and observed classification probabilities was +.72. For the independent cue model, the best rank-order correlation was +.41, considerably lower than that for the context model. The problem for the independent cue model is that A training stimuli were better classified than B training stimuli, a difference not predicted by the model. The parameters associated with the classification data were $c = .18, f = .20, s = .28$, and $n = .33$ for the context model and $W_c = .22$, $W_f = .27$, $W_s = .30$, $W_n = .30$, and $M_i = .54$ for the independent cue model. Table A6 in the Appendix summarizes these rank-order predictions.

Discussion

As predicted by the context model, the probability of calling a probe stimulus old increased with the number of training stimuli highly similar to the probe. In addition, the classification data were predicted considerably better by the context model than by the independent cue model.

General Discussion

Major Results

In each experiment, the data were more consistent with the context model than with the general independent cue model. Qualitative predictions favored the context model regardless of whether or not the two classes were linearly separable and regardless of whether the stimuli were geometric forms or face outlines. Moreover, quantitative predictions of the context model were in each case more accurate than corresponding predictions of the independent cue model, though generally these differences were small. At the very least, the context model must be taken seriously as a contending classification theory. The results also raise questions concerning the adequacy of the assumption that component stimulus dimensions are treated independently, a basic underlying assumption of prototype models in particular and independent cue theories in general.

The context model differs from all independent cue models in its assumption that

component dimensions are not independent and differs from all but average distance models in its proposal that classification judgments derive from exemplar, rather than category level, information. The model assumes that classification judgments are based on the retrieval of specific item information, the retrieval of information about old stimuli when a new stimulus is presented being a function of similarity. The model contains parameters for the similarity of the values along each dimension but adds no special process to differentiate old and new items. Nonetheless, the context model predicts differences between old and new stimuli, and by representing forgetting in terms of increases in the similarity parameters, it yields predictions consistent with the differential retention of old and new stimulus classification. Specific details of the various experiments were consistent with the context model, and an initial attempt at a quantitative fit to the data produced fairly accurate predictions. Finally, the new–old recognition data and the confidence rating data are consistent with the idea that recognition and confidence are related to how easily or quickly the probe stimulus retrieves specific stored information.

Relation of the Context Model to Other Theories

Although transfer performance is not well predicted by the frequency of individual features (Franks & Bransford, 1971), a number of researchers have proposed that subjects may encode relational frequencies (Hayes-Roth & Hayes-Roth, 1977; Neumann, 1974, 1977; Reitman & Bower, 1973). That is, in addition to encoding the frequency with which Features A, B, and C occur, subjects may encode the frequency with which A and B; B and C; A and C; and A, B, and C occur together. Judgments are assumed to be based on combinations of simple frequency (A, B, and C) and relational frequency (AB, AC, BC, and ABC). In most experiments, as distance of an exemplar from the prototype increases, relational frequency decreases; therefore, relational frequency models can account for most results concerning transfer to new patterns. Relational frequency theory and prototype theory predict transfer about equally well when geometric stimuli are used (Neumann, 1974; Posnansky & Neumann, 1976), while relational frequency models actually are better predictors of transfer performance than prototype theory when letter strings and biographical descriptions are used (Hayes-Roth & Hayes-Roth, 1977; Posnansky & Neumann, 1976). Whether these differences in accuracy of prediction depend on the nature of the stimuli presented or on details of experimental design that have covaried with the alternative stimuli is uncertain.

Relational frequency models differ from the context model in that they assume that category judgments are based on category level rather than exemplar information, but like the context model, they do not assume that category judgments are based on an independent summation of component information. In fact, most of the qualitative predictions of the context model examined in the present experiments are shared by relational frequency models. Several difficulties arise in attempting to extend these models to the present data with greater explicitness. For Neumann's attribute frequency model, the net frequency scores of old and new stimuli in our experiments frequently overlapped considerably, while ratings of new stimuli generally fell below those of old stimuli. Therefore, one would have to posit that a mixture of specific item information and category level (frequency) information controlled performance. A more serious problem concerns how to treat differential salience of dimensions. If color is more salient than form, one might imagine that a color frequency counter is more likely to be incremented than a simple form frequency counter, but how should one treat the status of the counter corresponding to color plus form? Another concern is that the encoding processes seem quite complex. When a pattern having values along four dimensions is presented, four one-dimensional counters, six two-dimensional counters, four three-dimensional counters, and one four-dimensional counter all get incremented and associated with the category. Probably none of these difficulties is insurmountable, but at present, the context model appears to be easier to work with, if not more parsimonious.

Hayes-Roth and Hayes-Roth (1977) have proposed a model called the *most diagnostic*

property set model, where classification is based on the features or combinations of features that are most diagnostic (i.e., have the highest validity). The fact that combinations of features may be used distinguishes this model from independent cue theories. As developed by Hayes-Roth and Hayes-Roth (1977), the property set model makes no differential predictions among stimuli that have at least one perfectly valid cue. In our experiments, the training stimuli always contain at least one perfectly valid combination of features, and transfer stimuli almost always contain at least one combination of features valid for each of the two categories. As a result, the Hayes-Roth and Hayes-Roth model makes almost no specific predictions for our experiments, except that performance on old patterns should be better than performance on new patterns. If one assumed that with diagnosticity held constant, classification performance varied with the number and frequency of the associated property sets, then the main qualitative results of our experiments would be predicted.

In the experiments designed to test the property set model, Hayes-Roth and Hayes-Roth (1977) varied the frequency with which the different exemplars were presented in training. Differential frequency of exemplar presentation might naturally be represented in the context model by assuming that different numbers of representations are developed. For example, if stimulus i appears twice as often as stimulus j, we might treat the task as involving two stimulus i exemplars and one stimulus j exemplar. With this interpretation, the context model would reproduce the main qualitative results of the Hayes-Roths' experiment.

Finally, one other very simple classification model, the proximity model, deserves consideration. One version of the proximity model simply assumes that new patterns are classified according to the training stimulus that most closely resembles the new stimulus. If that training stimulus was associated with Class A, then the new pattern is classified as an A. If the training stimulus was a B, then the new pattern would be called a B (for more extensive treatment of proximity models, see Reed, 1972; Sebestyen, 1962).

The proximity model is similar to the context model in that it assumes that classification is based on specific item information. In fact, the proximity model might be thought of as a special case of the context model where performance is based on the first pattern retrieved, and the first pattern retrieved is always the most similar training pattern. The proximity model would need further elaboration to account for the differential retention of old and new stimulus patterns observed in Experiment 1. Modification with this end in mind would further increase the similarity between the context model and the proximity model.

Relation of the Context Model to Other Classification Results

Many of the results on prototype abstraction and classification with ill-defined rules derive from studies employing dot patterns. Although it is somewhat risky to discuss the application of the context model to these results so long as we are unable to specify the component dimensions, the context model seems qualitatively consistent with some of the main findings.

First of all, transfer to new stimulus patterns appears to increase with the number of exemplars comprising a category (e.g., Homa & Chambliss, 1975). This result could be derived from the context model by proposing that the greater the number of training patterns, the more likely it is that a new pattern will be highly similar to at least one training pattern from the correct category. If this were true, the average dissimilarity (or distortion level) of the training pattern would enter as a determiner of transfer in the following manner: For training patterns that were low distortions (quite similar to one another), increasing the number of exemplars per category would not facilitate transfer as much as when the training patterns were more varied. In other words, if training patterns were low distortions, no one of them might be similar enough to a new probe to be activated by it, so that one might expect an interaction between category size and level of distortion of category members. Exactly this interaction was obtained by Homa and Vosburgh (1976).

According to prototype theories, learning and transfer behavior depend on the distance

of alternative prototypes from each other and on the average distance of exemplar stimuli from the prototypes. Performance should be unaffected by how these distances arise. Barresi, Robbins, and Shain (1975) held overall average distance of dot patterns from the prototypes constant and varied whether each individual dot in a given dot pattern was about the same distance from the corresponding dot in the prototype from which it was derived (low-variance condition) or whether some individual dots were close and others far from the corresponding dot locations in the prototype. Since overall average distance is constant, prototype theory predicts no effect of variance. The results showed, however, that high-variance classes were learned faster and had better generalization than low-variance classes.

The context model predicts that a new item highly similar to one member but dissimilar from another member of a category would be assigned to that category with greater ease than a new item moderately similar to both category members. Likewise, in distinguishing two items in different categories on the basis of the values along two dimensions, one similar value and one distinct value should combine to be more discriminable than two values intermediate in distinctiveness. Thus, if one is willing to assume that physical and psychological distances were roughly equivalent in the Barresi et al. experiment, the context model correctly predicts that high-variance categories will be more discriminable than low-variance categories.

Generality

While we have indicated informally how the context model might apply to other classification studies, the generality of the context model to different stimulus populations and experimental procedures is an open question. One reason for optimism concerning generality is that basically the same ideas give an excellent account of a wide range of discrimination learning and transfer phenomena across a variety of subject populations (Medin, 1975).

Stimulus dimensions. Immediate extensions to other stimulus populations face some nontrivial obstacles. The problems center around developing appropriate descriptions of the features or dimensions in terms of which stimulus information is stored in memory. In the absence of such analyses, the experimenters' and subjects' representations of the stimuli may not coincide, relational cues (e.g., ratio of nose length to mouth width) may operate, and interactions caused by factors such as asymmetrical similarity (e.g., Atkinson & Estes, 1963; Bush & Mosteller, 1951; Tversky, 1977) may go undetected. Yet, many of the comparisons that one is most interested in making involve natural concepts whose component features are unknown and undoubtedly complex.

Certain tests of the context model are even possible in working with complex stimuli, such as human faces, whose constituent dimensions cannot be specified. Such tests would employ a confusion matrix (which could be independently derived) for the exemplars to be used in the classification task. The work of Shepard and Chang (1963) suggests that classification performance can be predicted from an identification confusion matrix so long as selective attention is minimized. The context model would predict that with between-category confusability held constant, classification performance on an exemplar would be better the greater within-category confusability. The basis for this prediction is that if an exemplar is highly confusable with other exemplars in its category, then when the exemplar is presented as a probe, it will lead to a correct categorization if either the information associated with the stored representation of that exemplar is retrieved or if information associated with the representations of the other confusable exemplars is accessed. If an exemplar is not confusable with (similar to) any of its fellow category exemplars, the latter source of correct classifications will not be available.

Test situations. There is reason to think that the context model will account for classification performances in settings where analytical strategies are either absent or irrelevant to the concept in question. Brooks (Note 2) trained subjects on a paired-associate task, where the stimuli were complex symbols and the responses were either animal or city names. Then, subjects were told that the concept "old world" versus "new world" (animals or cities)

was relevant, and they were asked to classify new complex symbols as old or new world on the basis of what they had previously learned. The results suggested that subjects classified new patterns by analogy with similar learned instances rather than by applying analytical strategies. Incidental concept learning proved to be more effective than explicit concept training in situations where the stimuli were complex and presumably difficult to analyze or where the conceptual tasks were disguised during learning.

Although it was developed without the benefit of any knowledge of Brooks's research, the retrieval assumptions of the context model closely embody Brooks's propositions concerning analogical thinking. Therefore, the context model is likely to be at least qualitatively successful in addressing incidental concept learning.

The idea that judgments may be based on retrieval of examples has also been applied to situations distinctly different from concept learning situations. Tversky and Kahneman (1973) proposed that people often evaluate the frequency or likelihood of events by the ease with which relevant instances come to mind. This strategy can be very efficient, but it can lead to systematic biases. For example, people can accurately estimate how many words they can recall from a category within some time limit, but they make gross errors on questions such as whether the letter *k* is more likely to be the first than the third letter of English words. Most subjects guessed that *k* is more likely to appear in the first position, perhaps because people can generate words beginning with *k* more easily than words which have *k* as the third letter.

There is also some evidence that retrieval of instances may play a role in semantic memory tasks. Holyoak and Glass (1975) reported evidence consistent with the idea that sentences such as "All birds are canaries" often are disconfirmed by retrieval of counterexamples (e.g., robins). They speculate that true judgment of such sentences as "All birds lay eggs" may be confirmed by induction after the retrieval of several positive instances (e.g., robins lay eggs, ducks lay eggs, and ostriches lay eggs).

A related implication of the context model is that classification judgments depend on contextual factors. For example, a rather strange looking four-legged animal may be much more likely to be classified as a dog when seen walking down the street on a leash than when seen running through the woods. Our assumption is similar to the principle of encoding specificity (Tulving & Thompson, 1971, 1973; Watkins, Ho, & Tulving, 1976) in that the presence of specific context plays a role in whether the specific stimulus information is accessed.

Further Considerations and Conclusions

If one were going to design a system to learn about the structural relations between stimulus classes and events (i.e., a natural concept learning device), what kind of system should one build? First of all, it would seem that one would be judicious in one's use of analytical processes, particularly if those processes accurately reflected concept identification models that focused on valid cues. One would not want a system that only learned exceptionless rules. Hence, in some sense, learning by analogy should be allowed for.

An analogical process would also serve the function of protecting the concept learning system from ignoring or throwing away information that might later prove critical when other concepts or categories were acquired. For example, in learning to tell the difference between dogs and cats, size is a good cue, whereas domesticity is not. But ignoring domesticity and learning only about size could prove costly later when one might want to distinguish dogs and wolves or dogs and coyotes.

At the present time, it is hard to make a perfectly general statement concerning how one ought to balance analytical and analogical processes. On the one hand, Reber (1967, 1969, 1976) has shown that active rule search can interfere with acquiring a grammar if that grammar involves very complex rules; on the other hand, a system that did not abstract the rules within its grasp might prove to be needlessly inefficient.

Finally, in designing a concept learning system, one might want to be guided by the simplicity of the proposed underlying proces-

ses. The context model, embodying instance learning and retrieval on the basis of similarity, appears to accomplish efficient mastery of ill-defined concepts with a minimum of processing machinery.

Reference Notes

1. Ratcliff, R. *A model of recognition memory.* Paper presented at the meeting of Mathematical Psychology, New York University, 1976.
2. Brooks, L. *Non-analytic concept formation and memory for instances.* Paper presented at the Social Science Research Council Conference on Human Categorization, Berkeley, Calif., 1976.

References

Atkinson, R. C., & Estes, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2). New York: Wiley, 1963.

Barresi, J., Robbins, D., & Shain, K. Role of distinctive features in the abstraction of related concepts. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, *1*, 360–368.

Bourne, L. E., Jr. Knowing and using concepts. *Psychological Review*, 1970, *77*, 546–556.

Bourne, L. E., Jr., & O'Banion, K. Memory for individual events in concept identification. *Psychonomic Science*, 1969, *16*, 101–103.

Bush, R. R., & Mosteller, F. A model for stimulus generalization and discrimination. *Psychological Review*, 1951, *58*, 413–423.

Calfee, R. C. Recall and recognition memory in concept identification. *Journal of Experimental Psychology*, 1969, *81*, 436–440.

Estes, W. K. An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory.* Washington, D.C.: Winston, 1972.

Estes, W. K. Memory and conditioning. In F. J. McGuigan & D. B. Lumsden (Eds.), *Contemporary approaches to conditioning and learning.* New York: Wiley, 1973.

Estes, W. K. Structural aspects of associative models for memory. In C. N. Cofer (Ed.), *The structure of human memory.* New York: W. H. Freeman, 1976.

Franks, J. J., & Bransford, J. D. Abstraction of visual patterns. *Journal of Experimental Psychology*, 1971, *90*, 65–74.

Goldman, D., & Homa, D. Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 375–385.

Hayes-Roth, B., & Hayes-Roth, F. Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 321–338.

Hirschfeld, S. L., Bart, W. M., & Hirschfeld, S. F. Visual abstraction in children and adults. *Journal of Genetic Psychology*, 1975, *126*, 69–81.

Holyoak, K. J., & Glass, A. L. The role of contradictions and counterexamples in the rejection of false sentences. *Journal of Verbal Learning and Verbal Behavior*, 1975, *14*, 215–239.

Homa, D., & Chambliss, D. The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, *1*, 351–359.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, 1973, *101*, 116–122.

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

Katz, J. J., & Postal, P. M. *An integrated theory of linguistic descriptions.* Cambridge, Mass.: MIT Press, 1964.

Lasky, R. E. The ability of six-year-olds, eight-year-olds, and adults to abstract visual patterns. *Child Development*, 1974, *45*, 626–632.

Levine, M. *A cognitive theory of learning: Research on hypothesis testing.* Hillsdale, N.J.: Erlbaum, 1975.

Medin, D. L. A theory of context in discrimination learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9). New York: Academic Press, 1975.

Medin, D. L. Theories of discrimination learning and learning set. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes.* Hillsdale, N.J.: Erlbaum, 1976.

Neumann, P. G. An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, 1974, *2*, 241–248.

Neumann, P. G. Visual prototype information with discontinuous representation of dimensions of variability. *Memory & Cognition*, 1977, *5*, 187–197.

Peterson, M. J., Meagher, R. B., Jr., Chait, H., & Gillie, S. The abstraction and generalization of dot patterns. *Cognitive Psychology*, 1973, *4*, 378–398.

Posnansky, C. J., & Neumann, P. G. The abstraction of visual prototypes by children. *Journal of Experimental Child Psychology*, 1976, *21*, 367–379.

Posner, M. I., & Keele, S. W. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 1968, *77*, 353–363.

Posner, M. I., & Keele, S. W. Retention of abstract ideas. *Journal of Experimental Psychology*, 1970, *83*, 304–308.

Reber, A. S. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 1967, *6*, 855–863.

Reber, A. S. Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology*, 1969, *81*, 115–119.

Reber, A. S. Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory*, 1976, *2*, 88–94.

Reed, S. K. Pattern recognition and categorization. *Cognitive Psychology*, 1972, *3*, 382–407.

Reed, S. K., & Friedman, M. P. Perceptual vs. conceptual categorization. *Memory & Cognition*, 1973, *1*, 157–163.

Reitman, J. S., & Bower, G. H. Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, 1973, *4*, 194–206.

Rosch, E. On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press, 1973.

Rosch, E. Cognitive reference points. *Cognitive Psychology*, 1975, *7*, 532–547. (a)

Rosch, E. Cognitive representation of semantic categories. *Journal of Experimental Psychology: General*, 1975, *104*, 192–233. (b)

Rosch, E. The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *1*, 303–322. (c)

Rosch, E., & Mervis, C. B. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 1975, *7*, 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology*, 1976, *8*, 382–439.

Sebestyen, G. S. *Decision-making processes in pattern recognition*. New York: Macmillan, 1962.

Shepard, R. N., & Chang, J. J. Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, 1963, *65*, 94–102.

Smith, E. E., Shoben, E. J., & Rips, L. J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 1974, *81*, 214–241.

Spence, K. W. The nature of discrimination learning in animals. *Psychological Review*, 1936, *43*, 427–449.

Strange, W., Keeney, T., Kessel, F. S., & Jenkins, J. J. Abstraction over time of prototypes from distractions of random dot patterns: A replication. *Journal of Experimental Psychology*, 1970, *83*, 508–510.

Sutherland, N. S., & Mackintosh, N. J. *Mechanisms of animal discrimination learning*. New York: Academic Press, 1971.

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

Tulving, E., & Thomson, D. S. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology*, 1971, *87*, 116–124.

Tulving, E., & Thomson, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 1973, *80*, 352–373.

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

Tversky, A., & Kahneman, D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 1973, *5*, 207–232.

Watkins, M. J., Ho, E., & Tulving, E. Context effects in recognition memory for faces. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 505–517.

Appendix

Table A1
Predicted and Observed Rank Error Data from Experiment 2 for the Context Model and the Independent Cue Model

Stimulus number	Observed rank	Predicted rank: Context model	Predicted rank: Independent cue model
4	7	7	3.5
7	3	3	5
15	2	1.5	2
13	6	5.5	6
5	5	5.5	7
12	9	9	8
2	8	8	9
14	4	4	3.5
10	1	1.5	1

Note. The rankings are arranged from least errors to most errors.

Table A2
Predicted and Observed Ranked Classification Data from Experiment 2 for the Context Model and the Independent Cue Model

Stimulus number	Observed rank	Predicted rank: Context model	Predicted rank: Independent cue model
4	11	9	4.5
7	4	3.5	8
15	9.5	1.5	2
13	4	7.5	6
5	9.5	7.5	9
12	7	11	10
2	7	10	11
14	4	5	4.5
10	1	1.5	1
1A	15	12	15
3B	12	15	12
6A	2	3.5	3
8B	13	13	13
9A	16	16	16
11A	14	14	14
16B	7	6	7

Note. The rankings are arranged from highest classification scores to lowest.

Table A3
Predicted and Observed Rank Error Data from Experiment 3 for the Context Model and the Independent Cue Model

Stimulus number	Observed rank	Predicted rank: Context model	Predicted rank: Independent cue model
4	4	5	4
5	6	6	7.5
7	2	3	4
13	7	7	6
15	1	1	1.5
2	8	9	9
10	3	2	1.5
12	9	8	7.5
14	5	4	4

Note. The rankings are ordered from least to most errors.

Table A4
Predicted and Observed Ranked Classification Data from Experiment 3 for the Context Model and the Independent Cue Model

Stimulus number	Observed rank	Predicted rank: Context model	Predicted rank: Independent cue model
4	3	6	4.5
7	3	2.5	8
15	6	2.5	3
13	8	10	7
5	11	12	9
12	13	13	13
2	11	11	11
14	3	2.5	4.5
10	5	2.5	1
1A	11	8.5	14
3B	14	15	12
6A	1	5	2
8B	9	8.5	10
9A	16	16	16
11A	15	14	15
16B	7	7	6

Note. The rankings are ordered from highest to lowest classification scores.

Table A5
Predicted and Observed Ranked Error and Reaction Time Data from the Overtraining Phase of Experiment 3 for the Context and Independent Cue Models

Stimulus number	Observed rank		Predicted rank	
	Latency	Error	Context model	Independent cue model
4	4	5	5	4
7	3	2	2	4
15	1	1	2	1.5
13	6	6	6	6
5	8	7	8	7.5
12	9	9	9	7.5
2	7	8	7	9
14	4	5	4	4
10	2	3	2	1.5

Note. Rankings are arranged according to increasing errors and latency.

Table A6
Predicted and Observed Ranked Classification Data from Experiment 4 for the Context Model and the Independent Cue Model

Stimulus number	Observed rank	Predicted rank	
		Context model	Independent cue model
2	4.5	7	10.5
4	7	1	8.0
5	2	4	1.5
7	3	10.5	12.4
13	10	5	8.0
15	4.5	6	4.5
1	11	12	10.5
3	13.5	8.5	1.5
8	9	8.5	8.0
11	7	10.5	14
16	12	13	4.5
6A	1	2	4.5
9A	16	16	15
10B	7	3	4.5
12A	15	14	15
14B	13.5	15	12.5

Note. The rankings are ordered from highest to lowest classification scores.

Received October 20, 1977