Stimulus Categorization

F. Gregory Ashby
W. Todd Maddox

The bacterium *E. coli* tumbles randomly in a molecular sea. When it encounters a stream of molecules that it categorizes as a nutrient, it suppresses tumbling and swims upstream to the nutrient source. A recently inseminated female mouse sniffs urine near her nest. If she categorizes it as from an unfamiliar male mouse, implantation and pregnancy are prevented (Bruce, 1959; Parkes & Bruce, 1962). A man views a long sequence of portraits taken from high school yearbooks. Even though he graduated almost 50 years ago, he is remarkably accurate at deciding whether an arbitrary face belongs to the category of his own high school classmates (Bahrick, Bahrick, & Wittlinger, 1975). All organisms divide objects and events in the environment into separate classes or categories. If they did not, they would die and their species would become extinct. Therefore, categorization is among the most important decision tasks performed by organisms (Ashby & Lee, 1993).

Technically, a categorization or classification task is one in which there are more stimuli than responses. As a result, a number of stimuli are assigned the same response. In contrast, an identification task is one in which there is a unique response for every stimulus. For example, many humans are in the category “women” and many objects are in the category “bells,” but only one human is identified as “Hillary Clinton” and only one object is

*Measurement, Judgment, and Decision Making*

Copyright © 1996 by Academic Press. All rights of reproduction in any form reserved.

251
identified as "the Liberty Bell." Although the theories and basic phenomena associated with categorization and identification are similar, this chapter focuses on categorization.

A categorization task is one in which the subject assigns a stimulus to one of the relevant categories. Many other tasks require the subject to access stored category information but not to make a categorization judgment. For example, in a typicality rating task the subject sees a category exemplar (i.e., a stimulus belonging to the category) and rates how typical or representative of the category it is. Other experiments might ask the subject to recall all the exemplars of a particular category. Although these related paradigms provide valuable information about category representation, space limitations prevent us from considering them in detail. Instead, we will focus on the standard categorization experiment.

Another important distinction is between categories and concepts. Although these terms are sometimes used interchangeably, we define a category as a collection of objects belonging to the same group and a concept as a collection of related ideas. For example, trees form a category and the many alternative types of love form a concept. When Ann Landers tells a reader that he is in lust rather than in love, she is doing something very similar to categorization. Many of the categorization theories discussed here make definite predictions about the cognitive processes required for such a judgment. Even so, the representations of categories and concepts are probably quite different and a discussion of the two is beyond the scope of this chapter.

I. THE CATEGORIZATION EXPERIMENT

This discussion may leave the impression that the focus of this chapter is narrow. However, the standard categorization experiment has many degrees of freedom, which can result in a huge variety of tasks. Some prominent options available to the researcher designing a categorization task are listed in Table 1. The first choice is the type of stimuli selected. One can choose stimuli that vary continuously along the relevant stimulus dimensions or that only take on some number of discrete values. The most limiting case is binary-valued dimensions. In many such experiments the two levels are "presence" and "absence." Several categorization theories make specific predictions only in the special case where the stimulus dimensions are binary valued. This is ironic because in natural settings binary-valued stimulus dimensions are rare, if they exist at all. For example, in one popular experimental paradigm that consistently uses binary-valued dimensions, subjects learn that a patient received a battery of medical tests, that the outcome of each test is either positive or negative, and that a certain pattern of test results is characteristic of a particular disease. The subjects then
### TABLE 1 Options in the Design of a Categorization Experiment

<table>
<thead>
<tr>
<th>Experimental components</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus dimensions</td>
<td>Continuous vs. discrete vs. binary valued</td>
</tr>
<tr>
<td></td>
<td>Separable vs. integral</td>
</tr>
<tr>
<td>Category structure</td>
<td>Overlapping vs. nonoverlapping</td>
</tr>
<tr>
<td></td>
<td>Few vs. many exemplars</td>
</tr>
<tr>
<td></td>
<td>Linearly vs. nonlinearly separable</td>
</tr>
<tr>
<td></td>
<td>Normal vs. nonnormal category distributions</td>
</tr>
<tr>
<td>Instructions to subjects</td>
<td>Supervised vs. unsupervised vs. free sorting</td>
</tr>
<tr>
<td></td>
<td>Well defined vs. partially defined vs. undefined categories</td>
</tr>
</tbody>
</table>

discover the outcome of a set of tests and make a diagnosis. Is this realistic? How many medical tests give binary-valued results? For example, high blood pressure could indicate heart disease, but blood pressure does not have either a single high value or a single low value. Instead, it is continuous valued. A physician might decide on the basis of some continuous-valued blood pressure level that a patient has high blood pressure, but then it is the decision that is binary valued, not the percept. Even a simple home pregnancy test is not binary valued. For a variety of reasons, the testing material will display a continuum of hues, even if the woman is pregnant.

When selecting stimuli, a second choice is about the interaction between pairs of stimulus dimensions. If the dimensions are separable it is easy to attend to one and ignore the other, whereas if they are integral it is either difficult or impossible to do so (e.g., Ashby & Maddox, 1994; Ashby & Townsend, 1986; Garner, 1974; Maddox, 1992). Prototypical separable dimensions are hue and shape and prototypical integral dimensions are hue and brightness (e.g., Garner, 1977; Garner & Felfoldy, 1970; Hyman & Well, 1968).

Another set of options concerns the construction of the contrasting categories. For example, they can be overlapping or nonoverlapping. Overlapping categories have at least one stimulus that is sometimes a member of one category and sometimes a member of another (also called probabilistic categories). Thus, whereas perfect performance is possible with nonoverlapping categories, if the categories are overlapping, even the optimal classifier will make errors. Although much of the empirical work in categorization has used nonoverlapping categories, many natural categories are overlapping. For example, a person might look like the prototype of one ethnicity but be a member of another. Overlapping categories are also theoretically important because they provide a strong test of categorization theories. Virtually all theories of categorization can account for the kind of error-free performance that occurs when subjects categorize typewritten characters as x's or
o’s (nonoverlapping categories), but only a few (if any) can account for the
errors that occur when subjects try to categorize handwritten characters as
c’s or a’s (overlapping categories).

When designing the contrasting categories, the experimenter must also
decide how many exemplars to place in each category. This factor may have
a crucial effect on the strategy the subject uses to solve the categorization
problem. With only a few exemplars in each category the subject literally
could memorize the correct response to every stimulus, but if the categories
contain many exemplars, the subject might be forced to use a more efficient
strategy.

The experimenter must also decide where to place the category exemplars
in the stimulus space. Many choices are possible. Often, exemplars are
positioned so that a particular decision rule maximizes categorization accu-
racy. For example, the exemplars might be positioned so that a dimensional
rule is optimal (i.e., ignore all stimulus dimensions but one). One important
choice, which selects between broad classes of rules, is whether to make a
pair of categories linearly or nonlinearly separable. If the categories are lin-
early separable, then categorization accuracy is maximized by a rule that
compares a linear combination of the stimulus dimensional values to some
fixed criterion value. One response is given if the weighted sum exceeds the
criterion and the other response is given if it does not. If a pair of categories
is nonlinearly separable, then no such linear combination exists. The dis-
tinction between linearly and nonlinearly separable categories is important
because several theories predict that linearly separable categories should be
significantly easier to categorize than nonlinearly separable categories. An-
other prominent solution to the problem of how to position the exemplars
in the stimulus space is to allow their position to be normally distributed on
each stimulus dimension (Ashby & Gott, 1988; Ashby & Maddox, 1990,
normally distributed exemplars, the categories are always linearly or qua-
dratically separable.

After the stimuli are selected and the categories are constructed, the
experimenter must decide what instructions and feedback to give the sub-
ject. In a supervised task, the subject is told the correct response at the end of
each trial. In an unsupervised task, no feedback is given after each response,
but the subject is told at the beginning of the experiment how many catego-
ries are relevant. Finally, in a free sorting (or clustering) task, there is no trial-
by-trial feedback and the subject is given no information about the number
of relevant categories. Instead, the subject is told to form his or her own
categories, using as many as seem required.

A task can be supervised only if an objectively correct response can be
identified on every trial. In such a case, we say that the categories are well
defined. Of course, well-defined categories may also be used in an unsupervised or free sorting task, but sometimes these tasks are run with categories in which no objectively correct response exists on any trial. Such undefined categories are quite common. For example, suppose a subject is shown color patches of varying hue and is asked to categorize them according to whether they make the subject feel happy or sad. Because the subject's affective state is unobservable, there is no way to decide which response is correct. Thus, this experiment is an example of unsupervised categorization with undefined categories. (It is not free sorting because the subject was told that the only possible categories are happy and sad.) Finally, partially defined categories are those for which a correct response is identified on some but not on all trials. The most common use of partially defined categories is in experiments that use training and transfer conditions. During the training phase, feedback is given on every trial, but during the transfer phase, no feedback is given. If a stimulus is presented for the first time during the transfer phase, it therefore will usually have no objectively correct response.

II. CATEGORIZATION THEORIES

Categorization theories come in many different types and are expressed in different languages. This makes them difficult to compare. In spite of their large differences, however, they all make assumptions about (1) representation, (2) category access, and (3) response selection. The representation assumptions describe the perceptual and cognitive representation of the stimulus and the exemplars of the contrasting categories. The response selection assumptions describe how the subject selects a response after the relevant information has been collected and the requisite computations have been performed.

The category access assumptions delineate the various categorization theories. These assumptions describe the information that must be collected from the stored category representations and the computations that must be performed on this information before a response can be made. At least five different kinds of theories have been popular. The classical theory assumes that a category can be represented as a set of necessary and sufficient conditions, so categorization is a process of testing whether the stimulus possesses each of these conditions (e.g., Bruner, Goodnow, & Austin, 1956; 1

1 Our use of the term well defined is different from that of Neisser (1967), who distinguished between well- and ill-defined categories. According to Neisser, well-defined categories are structured according to simple logical rules, whereas ill-defined categories are not. These definitions are somewhat ambiguous because the term simple is not rigorously defined. Rules that are easily verbalized are usually called simple, but, for example, it is unclear whether a rule that can be verbalized, but not easily, is also simple.
Smith & Medin, 1981). **Prototype theory** assumes that the category representation is dominated by the prototype, or most typical member, and that categorization is a process of comparing the similarity of the stimulus to the prototype of each relevant category (Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1977). **Feature-frequency theory** assumes the category representation is a list of the features contained in all exemplars of the category, along with their relative frequency of occurrence (Estes, 1986a; Franks & Bransford, 1971; Reed, 1972). Categorization is a process of analyzing the stimulus into its component features and computing the likelihood that this particular combination of features was generated from each of the relevant categories. **Exemplar theory** assumes the subject computes the similarity of the stimulus to each stored exemplar of all relevant categories and selects a response on the basis of these similarity computations (Brooks, 1978; Estes, 1986a; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). Finally, **decision bound theory** (also called general recognition theory) assumes the subject constructs a decision bound that partitions the perceptual space into response regions (not necessarily contiguous), one for each relevant category. On each trial, the subject determines the region in which the stimulus representation falls, and then emits the associated response (Ashby & Gott, 1988; Ashby & Lee, 1991, 1992; Ashby & Townsend, 1986; Maddox & Ashby, 1993).

Much of the work on testing and comparing these theories has focused on response accuracy. This is a good dependent variable because it has high ecological validity and is easy to estimate. On the other hand, response accuracy is a fairly crude, global measure of performance. In the language of Marr (1982), response accuracy is good at testing between models written at the computational level, but it is poor at discriminating between models written at the algorithmic level. This focus on response accuracy has not yet been a serious problem because the most popular categorization models are computational rather than algorithmic. That is, they specify what is computed, but they do not specify the algorithms that perform those computations.

Currently, however, there is an awakening interest in algorithmic level descriptions of the categorization process. A test between algorithmic level models often requires a dependent variable more sensitive than overall accuracy to the microstructure of the data. In the categorization literature, algorithmic level models are most frequently tested against trial-by-trial learning data, although response times could also be used. The most popular architecture within which to implement the various algorithms that have been proposed has been the connectionist network and virtually all of the current network models instantiate some version of feature-frequency theory (e.g., Gluck & Bower, 1988) or exemplar theory (e.g., Estes, 1993, 1994; Kruschke, 1992). It is important to realize, however, that network versions
of classical, prototype, or decision bound theories could also be constructed. Thus, there is no such thing as the connectionist theory of categorization. Rather, connectionist networks should be viewed as an alternative architecture via which any computational theory of categorization can be expressed at the algorithmic level.

The next three sections of this chapter examine the representation, response selection, and category access assumptions in turn. This provides a common language from which to describe and formally compare the various theories. Section VI reviews the empirical tests of the various theories and the last section identifies some important unsolved problems.

III. STIMULUS, EXEMPLAR, AND CATEGORY REPRESENTATION

Figure 1 illustrates the relations between the various theories of stimulus and category representation. The most fundamental distinction is whether the theories assume numeric or nonnumeric representation. Nonnumeric models assume a symbolic or linguistic representation. These models assume that stimuli and category exemplars are described by a generative system of rules, which might be given by a production system or a grammar. Each category is associated with a unique set of rules, so categorization is equivalent to determining which set of rules generated the stimulus. Early proponents of nonnumeric representation in psychology were Allen Newell and Herbert Simon (e.g., Newell & Simon, 1972; see, also, Anderson, 1975; Klahr, Langley, & Neches, 1987).

Numeric representation is of two types. Dimensional theories assume a geometric representation that contains a small number of continuous-valued dimensions. The most widely known examples in psychology are multidimensional scaling (MDS; Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b; Torgerson, 1958; Young & Householder, 1938) and signal detection theory (Ashby & Townsend, 1986; Green & Swets, 1966). Feature theories assume

![Diagram](image)

**FIGURE 1** Hierarchical relations among theories of stimulus representation.
the stimulus can be represented as a set of features, where a feature is either present or absent and often has a nested relation that is naturally represented in a treelike connected graph. Perhaps the most notable feature models in psychology were developed by Amos Tversky (Corter & Tversky, 1986; Sartath & Tversky, 1977; Tversky, 1972, 1977). Although some argue that feature models are nonnumeric (e.g., Pao, 1989), we classify them as numeric because it is usually possible to depict feature representations geometrically by defining a feature as a binary-valued dimension. In this case, each dimension contributes only one bit of information (i.e., presence or absence), so the resulting perceptual space frequently has many dimensions. With binary-valued dimensions, the natural distance measure is the Hamming metric, defined as the number of features on which the two stimuli disagree. The dissimilarity measure proposed by Tversky (1977) in his feature contrast model, generalizes Hamming distance.

Assumptions about stimulus representation are impossible to test in isolation. At the very least, extra assumptions must be added about how the subject uses the representation. One obvious choice is to assume that representations that are close together are similar or related. A number of attempts to criticize numeric representation, and especially dimensional representation, have been based on this assumption. As we will see, however, a critical point of contention is how one should define “close together.”

Most formal theories of categorization assume a numeric representation. In a number of exemplar models, application is restricted to experiments that use binary-valued stimuli, so the stimulus and category representations are featural. This includes the context model (Medin & Schaffer, 1978), the array-similarity model (Estes, 1986a), and several network models (Estes, 1993, 1994; Gluck & Bower, 1988; Hurwitz, 1990). Most other formal models assume a dimensional representation. Most of these use a multidimensional scaling (MDS) representation that assumes (1) the stimuli and category exemplars are represented as points in a multidimensional space and (2) stimulus similarity decreases with the distance between the point representations. Exemplar models based on an MDS representation include the generalized context model (GCM: Nosofsky, 1986) and ALCOVE (Kruschke, 1992), whereas MDS-based prototype models include the fuzzy logical model of perception (FLMP; Massaro & Friedman, 1990) and the comparative distance model (Reed, 1972).

The assumption that similarity decreases with psychological distance is controversial. A distance-based perceptual representation is valid only if the psychological distances satisfy a set of distance axioms (Ashby & Perrin, 1988; Tversky, 1977). These include the triangle inequality, symmetry, minimality, and that all self-distances are equal. Unfortunately, there is abundant empirical evidence against these axioms. For example, Tversky (1977) reported that subjects rate the similarity of China to North Korea to be less
than the similarity of North Korea to China, an apparent violation of symmetry (see also Krumhansl, 1978). The triangle inequality holds if the psychological distance between stimuli i and j plus the distance between stimuli j and k is greater than or equal to the distance between stimuli i and k. Although the triangle inequality is difficult to test empirically, Tversky and Gati (1982) proposed an empirically testable axiom, called the corner inequality, that captures the spirit of the triangle inequality. Tversky and Gati (1982) tested the corner inequality and found consistent violations for stimuli constructed from separable dimensions.

It is important to note, however, that although these results are problematic for the assumption that similarity decreases with psychological distance, they are not necessarily problematic for the general notion of dimensional representation, or even for the point representation assumption of the MDS model. For example, Krumhansl (1978) argued that similarity depends not only on the distance between point representations in a low dimensional space, but also on the density of representations around each point. Her distance-density model can account for violations of symmetry but not for violations of the triangle inequality. Nosofsky (1991b) argued that many violations of the distance axioms are due to stimulus and response biases and not to violations of the MDS assumptions.

Decision bound theory assumes a dimensional representation, but one that is probabilistic rather than deterministic. A fundamental postulate of the theory is that there is trial-by-trial variability in the perceptual information obtained from every object or event (Ashby & Lee, 1993). Thus, a stimulus is represented as a multivariate probability distribution. The variability is assumed to come from many sources. First, physical stimuli are themselves intrinsically variable. For example, it is well known that the number of photons emitted by a light source of constant intensity and constant duration varies probabilistically from trial to trial (i.e., it has a Poisson distribution; Geisler, 1989; Wyszecki & Stiles, 1967). Second, there is perireceptor noise in all modalities. For example, in vision, the amount of light reflected off the cornea varies probabilistically from trial to trial. Third, there is spontaneous activity at all levels of the central nervous system that introduces more noise (e.g., Barlow, 1956, 1957; Robson, 1975).

One advantage of probabilistic representation is that decision bound theory is not constrained by any of the distance axioms. When stimuli are represented by probability distributions, a natural measure of similarity is distributional overlap (Ashby & Perrin, 1988). The distributional overlap similarity measure contains MDS Euclidean distance measures of similarity as a special case but, unlike the distance measures, is not constrained by the distance axioms (Ashby & Perrin, 1988; Perrin, 1992; Perrin & Ashby, 1991).

In a categorization task, the stimulus is usually available to the subject up to the time that a response is made. Thus, long-term memory has little or
no effect on the representation of the stimulus. In contrast, exemplars of the competing categories are not available, so a decision process requiring exemplar information must access the exemplar representations from memory. As a consequence, the representation of category exemplars is affected critically by the workings of memory. Therefore, unlike a theory of stimulus representation, a complete theory of exemplar representation must model the effects of memory. Nevertheless, most categorization theories represent stimuli and exemplars identically. Recently, a few attempts have been made to model the effects of memory on exemplar representation, but these have been mostly limited to simple models of trace-strength decay (e.g., Estes, 1994; Nosofsky, Kruschke, & McKinley, 1992). Clearly, more work is needed in this area.

Another area where more sophisticated modeling is needed is in category representation. In exemplar theory, a category is represented simply as the union or set of representations of all exemplars belonging to that category. Prototype theory assumes the category representation is dominated by the category prototype. Feature-frequency theory assumes the category representation is a list of all features found in the category exemplars. Classical theory assumes a category is represented by a set of necessary and sufficient conditions required for category membership. Thus, there is considerable disagreement among the theories about how much consolidation of the category representation is performed by the memory processes over time. Exemplar theory takes the extreme view that there is little or no consolidation, whereas classical theory posits so much consolidation that exemplar information is no longer available.

Although decision bound theory makes no concrete assumptions about category representation, several applications have tested the hypothesis that subjects assume categories are normally distributed (e.g., Ashby & Gott, 1988; Ashby & Lee, 1991; Ashby & Maddox, 1990, 1992; Maddox & Ashby, 1993). There is good evidence that many natural categories share properties of the normal distribution or at least that subjects assume that they do. First, natural categories generally contain a large number of exemplars (e.g., there are many trees). Second, the dimensions of many natural categories are continuous valued. Third, many natural categories overlap. Finally, there is evidence that people naturally assume category exemplars are unimodally and symmetrically distributed around some prototypical value (Fried & Holyoak, 1984; Flannagan, Fried, & Holyoak, 1986). As early as 1954, Black argued that "if we examine instances of the application of any biological term, we shall find ranges, not classes—specimens (i.e., individuals or species) arranged according to the degree of their variation from certain typical or 'clear' cases" (p. 28). The normal distribution has all these properties. It assumes an unlimited number of exemplars, dimensions that are
continuous valued, a small number of atypical exemplars (so it overlaps with other nearby categories), and it is unimodal and symmetric.

According to this interpretation, subjects initially assume the exemplars of an unfamiliar category have a multivariate normal distribution in stimulus space. Gaining experience with a category is a process of estimating the mean exemplar value on each dimension, the variances, and the correlations between dimensions. These estimates allow the subject to compute the likelihood that any stimulus belongs to this category. In fact, subjects need not even assume normality. Suppose they estimate the exemplar means, variances, correlations, and category base rates and then try to infer the correct distribution. If they do not know the appropriate family of distributions, it turns out that the multivariate normal is an excellent choice because it takes maximal advantage of the information available (technically, it is the maximum entropy inference; Myung, 1994). Given estimates of the category means, variances, correlations, and base rates, to infer that the category distribution is anything other than normal requires extra assumptions. In other words, the normal distribution is the appropriate noncommittal choice in such situations. Thus, the multivariate normal distribution is an attractive model of category representation (Ashby, 1992a; Fried & Holyoak, 1984; Flanagan, Fried, & Holyoak, 1986).

IV. RESPONSE SELECTION

There are two types of response selection models. Deterministic models assume that, if on different trials the subject receives the same perceptual information and accesses the same information from memory, then the subject will always select the same response. Probabilistic models assume the subject always guesses, although usually in a sophisticated fashion. In other words, if the evidence supports the hypothesis that the stimulus belongs to category A, then a deterministic model predicts that the subject will respond A with probability 1, whereas a probabilistic model predicts that response A will be given with probability less than 1 (but greater than 0.5).

In many categorization experiments, observable responding is not deterministic. It is not uncommon for a subject to give one response the first time a stimulus is shown and a different response the second time, even if the subject is experienced with the relevant categories (e.g., Estes, 1995). It is important to realize that such data do not necessarily falsify deterministic response selection models. The observable data may be probabilistic because of noise in the subject’s perceptual and memory systems. For example, perceptual noise may cause the subject to believe that a different stimulus was presented, a stimulus belonging to the incorrect category. Thus, the distinction between deterministic and probabilistic response selection models
does not apply at the observable level of the data but at the unobservable level of the subject's decision processes.

The question of whether response selection is deterministic or probabilistic is not limited to categorization tasks but may be asked about any task requiring an overt response from the subject. In many tasks, the evidence overwhelmingly supports deterministic response selection. For example, if subjects are asked whether individual rectangles are taller than they are wide or wider than they are tall, then, even at the data level, responding is almost perfectly deterministic (Ashby & Gott, 1988). In other tasks, the evidence overwhelmingly supports probabilistic response selection. In a typical probability matching task, a subject sits in front of two response keys. The right key is associated with a red light and the left key is associated with a green light. On each trial, one of these two lights is turned on. The red light is turned on with probability \( p \) and the green light is turned on with probability \( 1 - p \). The subject's task is to predict which light will come on by pressing the appropriate button. Consider the case in which \( p \) is considerably greater than one-half. A deterministic rule predicts that the subject will always press the right key. This choice also maximizes the subject's accuracy. However, the data clearly indicate that subjects sometimes press the right key and sometimes the left. In fact, they approximately match the objective stimulus presentation probabilities by pressing the right key on about 100\(p\%\) of the trials (e.g., Estes, 1976; Herrnstein, 1961, 1970). This behavior is known as probability matching. Therefore, the consensus is that humans use deterministic response selection rules in some tasks and probabilistic rules in other tasks. In categorization however, the controversy is still unresolved. It is even possible that subjects use deterministic rules in some categorization tasks and probabilistic rules in others (Estes, 1995).

Virtually all models assuming a probabilistic response selection rule, assume a rule of the same basic type. Consider a categorization task with categories \( A \) and \( B \). Let \( S_{ij} \) denote the strength of association between stimulus \( i \) and category \( J \) (\( J = A \) or \( B \)). The algorithm used to compute this strength will depend on the specific categorization theory. For example, in some prototype models, \( S_{iA} \) is the similarity between stimulus \( i \) and the category \( A \) prototype. In an exemplar model, \( S_{iA} \) is the sum of the similarities between the stimulus and all exemplars of category \( A \). In many connectionist models, \( S_{iA} \) is the sum of weights along paths between nodes activated by the stimulus and output nodes associated with category \( A \). Virtually all categorization models assuming a probabilistic response selection rule assume the probability of responding \( A \) on trials when stimulus \( i \) is presented equals

\[
P(R_A|i) = \frac{\beta_A S_{iA}}{\beta_A S_{iA} + \beta_B S_{iB}},
\]

(1)
where $\beta_j$ is the response bias toward category $j$ (with $\beta_j \geq 0$). Without loss of generality, one can assume that $\beta_B = 1 - \beta_A$. In many categorization models the response biases are set to $\beta_A = \beta_B = 0.5$.

Equation (1) has various names. It was originally proposed by Shepard (1957) and Luce (1963), so it is often called the Luce-Shepard choice model. But it is also called the similarity-choice model, the biased-choice model, or the relative-goodness rule. If $S_{iA}$ is interpreted as the evidence favoring category $A$, and if there is no response bias, then Eq. (1) is also equivalent to probability matching.

Deterministic decision rules are also of one basic type. Let $h(i)$ be some function of the stimulus representation with the property that stimulus $i$ is more likely to be a member of category $A$ when $h(i)$ is negative and a member of category $B$ when $h(i)$ is positive. For example, in prototype or exemplar models $h(i)$ might equal $S_{iB} - S_{iA}$. The deterministic decision rule is to

\[
\text{respond } A \text{ if } h(i) < \delta + e; \text{ respond } B \text{ if } h(i) > \delta + e. \tag{2}
\]

In the unlikely event that $h(i)$ exactly equals $\delta + e$, the subject is assumed to guess. As with the $\beta$ parameter in the similarity-choice model, $\delta$ is a response bias. Response $A$ is favored when $\delta > 0$ and response $B$ is favored when $\delta < 0$. The random variable $e$, represents criterial noise; that is, variability in the subject’s memory of the criterion $\delta$. It is assumed to have a mean of 0 and a variance of $\sigma^2$ and is usually assumed to be normally distributed (e.g., Maddox & Ashby, 1993).

Although the similarity-choice model and the deterministic decision rule of Eq. (2) appear very different, it is well known that the similarity-choice model is mathematically equivalent to a number of different deterministic decision rules (e.g., Marley, 1992; Townsend & Landon, 1982). Ashby and Maddox (1993) established another such equivalence that is especially useful when modeling categorization data. Suppose the subject uses the deterministic response selection rule of Eq. (2) and he or she defines the discriminant function $h(i)$ as $h(i) = \log(S_{iB}) - \log(S_{iA})$. Assume the criterial noise $e$ has a logistic distribution. Ashby and Maddox (1993) showed that under these conditions the probability of responding $A$ on trials when stimulus $i$ is presented is equal to

\[
P(R_A|i) = \frac{\beta_A(S_{iA})^\gamma}{\beta_A(S_{iA})^\gamma + \beta_B(S_{iB})^\gamma}, \tag{3}
\]

where

\[
\gamma = \frac{\pi}{\sqrt{3\sigma_e}} \quad \text{and} \quad \beta_A = \frac{e^\gamma}{1 - e^\gamma}.
\]
In other words, the probability matching behavior of the similarity-choice model, which results when \( \gamma = 1 \), is indistinguishable from a deterministic decision rule in which \( \sigma_e = \pi/\sqrt{3} \). On the other hand, if the subject uses a deterministic decision rule, but \( \sigma_e < \pi/\sqrt{3} \), then there is an equivalent probabilistic decision rule of the Eq. (3) type in which \( \gamma > 1 \). In this case, the observable responding is less variable than predicted by probability matching and the subject is said to be overmatching (Baum, 1974). Similarly, if \( \sigma_e > \pi/\sqrt{3} \), then there is an equivalent probabilistic decision rule in which \( \gamma < 1 \). In this case, the observable responding is more variable than predicted by probability matching and the subject is said to be undermatching (Baum, 1974).

These results indicate that for any deterministic response selection rule there is a probabilistic rule that is mathematically equivalent (and vice versa). Despite this fact, there is some hope for discriminating between these two strategies. This could be done by fitting the Eq. (3) model to categorization data from a wide variety of experiments and comparing the resulting estimates of \( \gamma \). For example, suppose a subject is using the deterministic rule of Eq. (2). If so, there is no good reason to expect \( \sigma_e \) to turn out to equal \( \pi/\sqrt{3} \) exactly (the value equivalent to probability matching). Also, it is reasonable to expect \( \sigma_e \) to vary with the nature of the stimuli and the complexity of the rule that separates the contrasting categories. Therefore, if the estimates of \( \gamma \) are consistently close to 1.0 or even if they are consistently close to any specific value, then a probabilistic rule is more likely than a deterministic rule. On the other hand, if the \( \gamma \) estimates vary across experiments and especially if they are larger in those tasks where less criterial noise is expected, then a deterministic rule is more likely than a probabilistic rule.

Another possibility for testing between deterministic and probabilistic response selection rules is to examine the \( \gamma \) estimates as a function of the subject's experience in the task. With probabilistic rules, \( \gamma \) might change with experience, but there is no reason to expect a consistent increase or decrease with experience. On the other hand, as the subject gains experience in the task, criterial noise should decrease because the strength of the subject's memory trace for the rule that separates the contrasting categories should increase. Thus, deterministic decision rules predict a consistent increase of \( \gamma \) with experience (Koh, 1993).

A number of studies explicitly tried to test whether subjects use deterministic or probabilistic response selection rules in a simple type of categorization experiment called the numerical decision task (Hammerton, 1970; Healy & Kubovy, 1977; Kubovy & Healy, 1977; Kubovy, Rappoport, & Tversky, 1971; Lee & Janke, 1964, 1965; Ward, 1973; Weissmann, Hollingsworth, & Baird, 1975). In these experiments, stimuli are numbers and
two categories are created by specifying two different normal distributions. On each trial, a number is sampled from one of the distributions and shown to the subject. The subject's task is to name the category (i.e., the distribution) from which it was drawn. In general, these studies have favored deterministic rules over probabilistic rules. For example, Kubovy et al. (1971) found that a fixed cutoff accounted for the data significantly better than a probability matching model, even when the probability matching model was allowed a response bias parameter.

Maddox and Ashby (1993) fit the Eq. (3) response selection model to data from 12 different categorization experiments. Category similarity $S_{ij}$ was computed from a powerful exemplar model. In several of these experiments the stimuli were rectangles. The category prototypes for two of these experiments are shown in Figure 2. In both cases the contrasting categories were linearly separable. In the first case, the rule that maximized accuracy was as follows:

Respond $A$ if the stimulus rectangle is higher than it is wide.
Respond $B$ if it is wider than it is high.

![Figure 2](image_url)

**Figure 2** Category prototypes for two categorization experiments.
In the second case, the optimal rule was as follows:

Respond A if the height plus the width is greater than some criterion amount.

Respond B if it is less than this criterion amount.

A major difference between these two tasks is that the second task requires the subject to maintain a criterion in memory, whereas the first task does not. Thus, if the subject is using a deterministic response selection rule there should be virtually no criterial noise in the first task but a significant amount in the second, and as a result \( \gamma \) should be much larger in the first task than the second. On the other hand, the Eq. (3) probabilistic rule must predict that \( \gamma \) is the same in the two tasks. The stimuli were the same, the instructions were the same, and optimal accuracy was the same. The categories even had the same amount of variability in the two tasks. As it turned out however, in the first task the median \( \gamma \) estimate was 2.59, whereas in the second task it was 1.00 (the median was taken across subjects). This is compelling evidence that subjects used a deterministic response selection rule in these tasks.

Estes (1995) argued that subjects may have used deterministic response selection rules in these tasks because "when stimuli are defined on only one or two sensory dimensions, subjects can discover a criterion that defines category membership (e.g., all angles greater than 45° belong to Category A) and recode stimuli in terms of their relation to the criterion, whereas with complex, multiattribute stimuli such recoding may be difficult or impossible" (p. 21). Several other data sets fit by Maddox and Ashby (1993) provide at least a partial test of this hypothesis. Six experiments used categories in which the optimal decision rule was highly nonlinear (it was quadratic). In at least four of these cases, there was no straightforward verbal description of this rule, so it would be extremely difficult for subjects to perform the kind of recoding that Estes describes. The subjects in these experiments all completed several experimental sessions and Maddox and Ashby (1993) fit the data of each individual session separately. Across the six experiments, the median \( \gamma \) estimates (computed across subjects) ranged from 1.13 to 4.29 on the first experimental session and from 1.51 to 5.67 on the last session. Thus, even when the subjects were inexperienced, observable responding was less variable than predicted by probability matching. More interesting however, is a comparison of the \( \gamma \) estimates from the first session to the last. In all six experiments, the estimated value of \( \gamma \) increased from the first session to the last. These results favor the deterministic response selection hypothesis. It is true however, that the stimuli in these six experiments varied on only two physical dimensions. Stimuli were rectangles that varied in height and width or circles that varied in size and orientation of a radial line. Thus, the Maddox and Ashby (1993) results do
The possibility that subjects switch to a probabilistic response when the stimuli vary on many dimensions.

Also examined the effects of experience on the amount of observable responding. Her stimuli were lines that varied in inclination and she used categories that were overlapping and able. As a measure of response variability, she estimated a $\tau$ is essentially equivalent to $\gamma$. For all subjects that were able to $\gamma$ increased with practice and eventually asymptoted at values larger than 1. Perhaps more interesting, however, was that $\delta$ was refit after the data were averaged across subjects, the slope of $\gamma$ was very close to 1. In other words, although almost were overmatching, the averaged data satisfied probability is apparent paradox occurred because, although each subject used some decision rule, different subjects settled on different here was no single rule that described the averaged data.

This indicates that experimental conditions can have a large effect on the resulting data. Support for deterministic or probabilistic selection rules. Some of the more important experimental results appear in Table 2. Responding will usually be less variable if the category is more accurately described by a simple verbal description, are highly practiced, and if single subject analyses are performed. Any factors that reduce perceptual or criterial noise should result in less variable. Perceptual noise can be reduced by using rather low contrast displays and response terminated stimulus presentation. Criterial noise can be reduced if the external rather than an internal criterion or referent. On the other hand, responding will usually be more variable if the optimal rule has an internal and external verbal description. Subjects are inexperienced in the data are averaged across subjects.

These experimental factors might affect the appearance of the no reason to believe that they will induce a subject to switch,

| Experimental Conditions Most Likely to Produce Data That Appear to Support Deterministic or Probabilistic Response Selection Rules |
|--------------------------------------------------|--|
| Simple response selection                        | Probabilistic response selection |
| hele has verbal analogue)                        | Optimal rule is complex          |
| Uses external criterion emory requirement)       | (optimal rule has no verbal analogue) |
| Subjects                                         | Optimal rule uses internal criterion |
| Analyses                                         | (extensive memory requirement)    |
|                                                  | Inexperienced subjects           |
|                                                  | Averaging across subjects        |
say, from a deterministic to a probabilistic response selection rule. Until there is good evidence to the contrary, the simplest hypothesis is that subjects use the same type of response selection rule in virtually all categorization tasks. Because of the identifiability problems, deterministic and probabilistic rules are difficult but not impossible to discriminate between. Currently, the best available evidence favors deterministic rules, but the debate is far from resolved.

V. CATEGORY ACCESS

This section reviews five major theories about the type of category information that is accessed and the computations that are performed on that information during categorization. Before beginning however, it is instructive to consider the optimal solution to the category access problem.

The optimal classifier uses the decision rule that maximizes categorization accuracy. Consider a task with two categories A and B. Suppose the stimulus is drawn from category A with probability \( P(C_A) \) and from category B with probability \( P(C_B) \). Let \( f_A(i) \) and \( f_B(i) \) be the likelihood that stimulus \( i \) is a member of category A or B, respectively. Then on trials when stimulus \( i \) is presented, the optimal classifier uses the deterministic rule:

\[
\begin{align*}
\text{if } & \frac{f_A(i)}{f_B(i)} > \frac{P(C_B)}{P(C_A)} \text{ then respond A } \\
\text{if } & \frac{f_A(i)}{f_B(i)} = \frac{P(C_B)}{P(C_A)} \text{ then guess } \\
\text{if } & \frac{f_A(i)}{f_B(i)} < \frac{P(C_B)}{P(C_A)} \text{ then respond B .}
\end{align*}
\]

The set of all stimuli for which \( f_A(i)/f_B(i) = P(C_B)/P(C_A) \) is a decision bound because it partitions the perceptual space into response regions. In general, the optimal decision bound can have any shape, but if each category representation is a multivariate normal distribution, then the optimal decision bound is always linear or quadratic.

A subject who would like to respond optimally in a categorization task must solve several problems. First, in a real experiment, the subject will have experience with only a limited sample of exemplars from the two categories. Therefore, even with perfect memory and an error-free perceptual system it is impossible to estimate perfectly the category likelihoods \( f_A(i) \) and \( f_B(i) \). At best, the subject could compute imperfect estimates of \( f_A(i) \) and \( f_B(i) \) (and also of the base rates) and use these in the optimal decision rule (Ashby & Alfonso-Reese, 1995). In this case, the subject’s decision bound will not agree with the optimal bound. Assuming the subject chooses this path, the next problem is to select an estimator.
In statistics, the likelihoods $f_A(i)$ and $f_B(i)$ are called probability density functions. Density function estimators are either parametric or nonparametric. Parametric estimators assume the unknown density function is of a specific type. In our language, this is equivalent to assuming some a priori category structure. For example, if the subject assumes that the category $A$ distribution is normal, then the best method of estimating $f_A(i)$ is to estimate separately the category mean and variance and insert these estimates into the equation that describes the bell-shaped normal density function (assuming only one relevant dimension). Nonparametric estimators make few a priori assumptions about category structure. The best known example is the familiar relative frequency histogram, but many far superior estimators have been discovered (e.g., Silverman, 1986). We will return to the idea of categorization as probability density function estimation later in this section.

Note that a subject who uses the optimal decision rule with estimates of $f_A(i)$ and $f_B(i)$ need not retrieve any exemplar information from memory. No matter what estimators are used, the updating required after each new stimulus could be done between trials. If it is, then when a new stimulus is presented the estimators would be intact and the two relevant likelihoods, that is, the estimates of $f_A(i)$ and $f_B(i)$ could be retrieved directly.

We turn now to an overview of the five major theories and then discuss the many empirical comparisons that have been conducted.

**A. Classical Theory**

The oldest theory of categorization is classical theory, which dates back to Aristotle, but in psychology was popularized by Hull (1920). Much of the recent work on classical theory has been conducted in psycholinguistics (Fodor, Bever, & Garrett, 1974; Miller & Johnson-Laird, 1976) and psychological studies of concept formation (e.g., Bourne, 1966; Bruner, Goodman, & Austin, 1956).

Classical theory makes unique assumptions about category representation and about category access. All applications of classical theory have assumed a deterministic response selection rule. First, the theory assumes that every category is represented as a set of singly necessary and jointly sufficient features (Smith & Medin, 1981). A feature is singly necessary if every member of the category contains that feature. For example, “four sides” is a singly necessary feature of the “square” category because every square has four sides. A set of features are jointly sufficient if any entity that contains the set of features is a member of the category. The features (1) four sides, (2) sides of equal length, (3) equal angles, and (4) closed figure are jointly sufficient to describe the “square” category, because every entity with these four attributes is a square.
The category access assumptions follow directly from the representation assumptions. When a stimulus is presented, the subject is assumed to retrieve the set of necessary and sufficient features associated with one of the contrasting categories. The stimulus is then tested to see whether it possesses exactly this set of features. If it does, the subject emits the response associated with that category. If it does not, the process is repeated with a different category.

Although classical theory accurately describes many categorization tasks (e.g., classifying squares versus triangles), the theory is associated with a number of predictions that are known to be false. First, classical theory excludes categories that are defined by disjunctive features, whereas subjects can learn tasks in which the optimal rule is disjunctive, such as the biconditional or exclusive-or problems (e.g., Bourne, 1970; Bruner, Goodnow, & Austin, 1956; Haygood & Bourne, 1965).

Second, it is difficult to list the defining features of many categories. For example, Wittgenstein (1953) argued that no set of necessary and sufficient features exists for the category “game.” Some games have a “winner” (e.g., football), but others do not (e.g., ring-around-the-rosie).

The third, and perhaps strongest, evidence against classical theory, is the finding that categories possess graded structure—that is, members of a category vary in how good an example (or how typical) they are of the category. Graded structure has been found in nearly every category (see Barsalou, 1987, for a review). For example, when asked to judge the typicality of different birds, subjects reliably rate the robin as very typical, the pigeon as moderately typical, and the ostrich as atypical (Mervis, Catlin, & Rosch, 1976; Rips, Shoben, & Smith, 1973; Rosch, 1973). In addition, if subjects are asked to verify whether a stimulus belongs to a particular category, response accuracy increases and response time decreases as typicality increases (although only on YES trials; e.g., Ashby, Boynton, & Lee, 1994; Rips et al., 1973; Rosch, 1973). Interesting typicality effects have also been found in the developmental literature. For example, typical category members are learned first by children (Rosch, 1973; Mervis, 1980) and are named first when children are asked to list members of a category (Mervis et al., 1976). Classical theory, on the other hand, predicts that all members of a category are treated equally, because they all share the same set of necessary and sufficient features.

Note that all three of these criticisms are directed at the category representation assumptions of classical theory, not at the category access assumptions. Thus, none of these results rule out the possibility that the category access assumptions of classical theory are basically correct. Especially relevant to this observation is the fact that most of the data on which the criticisms are based were not collected in categorization tasks (but rather, e.g., in typicality rating tasks). The major exception is the fact that subjects
can learn categorization tasks in which the optimal rule is disjunctive. The simplest way to handle this is to generalize the classical theory to allow a category to be defined as the union of subcategories, each of which is defined by a set of necessary and sufficient features (e.g., Ashby, 1992a, 1992b; Smith & Medin, 1981; see also, Huttenlocher & Hedges, 1994). For example, the category “games” could be defined as the union of “competitive games” and “noncompetitive games.”

A classical theorist could respond to the other criticisms by arguing that an exemplar-based graded category representation exists, and whereas this graded representation is used in recall and typicality rating tasks, it is not accessed on trials of a categorization task. Instead, when categorizing, the subject only needs to retrieve the categorization rule, which according to classical theory is a list of necessary and sufficient features for each relevant category or subcategory. This more sophisticated version of classical theory can only be falsified by data from categorization experiments. As it turns out, such data is not difficult to collect (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990), but as we will see, the notion that some category related tasks access a graded category representation whereas other such tasks do not is more difficult to disconfirm.

B. Prototype Theory

The abundant evidence that category representations have a graded structure led to the development of prototype theory (e.g., Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973; Rosch, Simpson, & Miller, 1976). Instead of representing a category as a set of necessary and sufficient features, prototype theory assumes that the category representation is dominated by the prototype, which is the most typical or ideal instance of the category. In its most extreme form, the prototype is the category representation, but in its weaker forms, the category representation includes information about other exemplars (Busemeyer, Dewey, & Medin, 1984; Homa, Dunbar, & Nohre, 1991; Shin & Nosofsky, 1992). In all versions however, the prototype dominates the category representation.

Much of the early work on prototype theory focused on recall and typicality rating experiments; that is, on tasks other than categorization. Two alternative prototype models have been developed for application to categorization tasks. The first, developed by Reed (1972), assumes a multidimensional scaling (MDS) representation of the stimuli and category prototypes. On each trial, the subject is assumed to compute the psychological distance between the stimulus and the prototype of each relevant category. Reed’s

---

2 More formally, the distribution of exemplars in the superordinate category is a probability mixture of the exemplars in the subordinate categories.
model assumed a deterministic response selection rule (i.e., respond with
the category that has the nearest prototype), but versions that assume prob-
abilistic response selection have also been proposed (e.g., Ashby & Maddox,
1993; Nosofsky, 1987; Shin & Nosofsky, 1992). We refer to all of these
as MDS-prototype models. The other prominent prototype model is called
the fuzzy-logical model of perception (FLMP, Cohen & Massaro, 1992;
Massaro & Friedman, 1990). The FLMP assumes a featural, rather than a
dimensional, representation of the stimuli and category prototypes. It also
assumes that a stimulus is compared to each prototype by computing the
fuzzy-truth value (Zadeh, 1965) of the proposition that the two patterns are
composed of exactly the same features.3 Recently, Crowther, Batchelder,
and Hu (1995) questioned whether the fuzzy-logical interpretation purportedly offered by the FLMP is warranted. Response selection in the FLMP is
probabilistic [the Eq. (1) similarity-choice model with all response biases set
equal]. Although the FLMP appears to be quite different from the MDS-
prototype model, Cohen and Massaro (1992) showed that the two models
make similar predictions.

Although prototype theory was seen as a clear improvement over classical
theory, it quickly began to suffer criticisms of its own. If the prototype is
the only item stored in memory, then all information about category vari-
ability and correlational structure is lost. Yet several lines of research sug-
gested that nonprototypical category exemplars can have a pronounced ef-
fect on categorization performance (e.g., Brooks, 1978; Hayes-Roth &
Hayes-Roth, 1977; Medin & Schaffer, 1978; Medin & Schwanenflugel,
1981; Neumann, 1974; Reber, 1976; Reber & Allen, 1978; Walker, 1975). In
particular, subjects are highly sensitive to the correlational structure of the
categories (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Me-
din, Aleom, Edelson, & Freko, 1982; Medin & Schwanenflugel, 1981;
Nosofsky, 1986, 1987, 1989). Note that this criticism is directed at the
category access assumptions of prototype theory, not at the category repre-
sentation assumptions.

Rosch (1975, 1978) understood the importance of the criticisms against
prototype theory and in an attempt to strengthen the theory argued that
almost all categories contain multiple prototypes. In fact, she argued that
“in only some artificial categories is there by definition a literal single proto-
type” (p. 40). For example, both robin and sparrow seem to be prototypes
for the category “bird.” In this spirit, Anderson (1990, 1991) proposed a
multiple prototype model, called the rational model. The rational model

3 The overall fuzzy-truth value of this proposition is equal to the product of the fuzzy-truth
values of the propositions that each specific feature of the stimulus is equal to the analogous
feature in the prototype (Massaro & Friedman, 1990). Fuzzy-truth value has many of the
properties of similarity, so the FLMP product rule is analogous to the assumption that simi-
arity is multiplicative across dimensions (Nosofsky, 1992b).
assumes that the category representation is a set of clusters of exemplars, each of which is dominated by a prototype. The probability that an exemplar is grouped into a particular cluster is determined by (1) the similarity of the exemplar to the cluster’s prototype and (2) a prior probability that is determined by the number of exemplars in each cluster and by the value of a coupling parameter. When presented with a stimulus to be categorized, the subject is assumed to compute the similarity between the stimulus and the prototype of each cluster and to select a response on the basis of these similarity computations. Few empirical tests of this model have been conducted (however, see Ahn & Medin, 1992; Nosofsky, 1991a).

C. Feature-Frequency Theory

Feature-frequency theory (Estes, 1986a; Franks & Bransford, 1971; Reed, 1972) has its roots in feature-analytic models of pattern perception, which assume that a visual stimulus is perceived as the set of its constituent features (Geyer & DeWald, 1973; Gibson, Oser, Schiff, & Smith, 1963; Townsend & Ashby, 1982). A key assumption of feature-analytic models is feature-sampling independence, which states that the probability of perceiving features \( f_a \) and \( f_b \) equals the probability of perceiving feature \( f_a \) times the probability of perceiving feature \( f_b \) (Townsend & Ashby, 1982; Townsend, Hu, & Ashby, 1981). In other words, feature-analytic models assume separate features are perceived independently. Townsend and Ashby (1982) found strong evidence against this assumption and, as a consequence, feature-analytic models of pattern perception are no longer popular.

Feature-frequency theories of categorization borrow their stimulus representation assumptions from the feature-analytic models of pattern perception. Suppose stimulus \( i \) is constructed from features \( f_1, f_2, \ldots, f_n \). Then the strength of association of stimulus \( i \) to category \( j \), denoted as before by \( S_{ij} \), is assumed to equal

\[
S_{ij} = \hat{P}(C_j)\hat{P}(i|C_j) = \hat{P}(C_j) \prod_{k=1}^{n} \hat{P}(f_k|C_j),
\]

where \( \hat{P}(C_j) \) is the subject’s estimate of the a priori probability that a random stimulus in the experiment is from category \( j \), \( \hat{P}(i|C_j) \) is an estimate of the probability (or likelihood) that the presented stimulus is from category \( j \), and \( \hat{P}(f_k|C_j) \) is an estimate of the probability (or likelihood) that feature \( f_k \) occurs in an exemplar from category \( j \). The latter equality holds because of the sampling independence assumption. Some feature-frequency models assume the probabilistic response selection rule of Eq. (1) (e.g., Estes, 1986a; Gluck & Bower, 1988), and some assume the deterministic rule of Eq. (2) (with \( h(i) = S_{ij} - S_{iA} \), e.g., Reed, 1972).
Feature-frequency theory can take on many forms depending on what assumptions are made about how the subject estimates the feature frequencies, that is, the $P(f_j | C_i)$. The original, and perhaps the most natural, interpretation is that the category $j$ representation is a list of the features occurring in all exemplars of category $j$, along with the relative frequency with which each feature occurred (Franks & Bransford, 1971; Reed, 1972). Another possibility is that the feature frequencies of the presented stimulus are estimated by doing a feature-by-feature comparison of the stimulus to the category prototypes. This type of feature-frequency model is equivalent to a prototype model that assumes feature-sampling independence. A third possibility is that the category $j$ representation is the set of all exemplars that belong to category $j$. To estimate $P(f_j | C_i)$ the subject scans the list of stored category $j$ exemplars and computes the proportion that contain feature $f_j$. This interpretation leads to a special case of exemplar theory (Estes, 1986a).

Gluck (1991) argued against the feature-frequency model offered by Gluck and Bower (1988; i.e., the adaptive network model) on the basis of its failure to account for the ability of subjects to learn nonlinearly separable categories. It is important to note that not all feature-frequency models are so constrained. For example, suppose all features are continuous-valued and that the subject assumes the values of each feature are normally distributed within each category, with a mean and variance that varies from category to category. To estimate $P(f_j | C_i)$, the subject first estimates the mean and variance of the values of feature $f_j$ within category $j$ and then inserts these estimates into the equation for the probability density function of a normal distribution. Then categories $A$ and $B$ are nonlinearly separable and the subject will learn these categories (i.e., respond with a nonlinear decision bound) if the following conditions hold: (1) all feature values are normally distributed, (2) the values of all feature pairs are statistically independent (so that feature sampling independence is valid), (3) the values of at least one feature have different variances in the two categories, and (4) the subject uses the response selection rule of the optimal classifier, i.e., Eq. (4).

Given the strong evidence against the feature-sampling independence assumption (e.g., Townsend & Ashby, 1982) and the fact that so many of the feature-frequency models are special cases of the more widely known categorization models, we will have little else to say about feature-frequency theory.

D. Exemplar Theory

Perhaps the most popular approach to dealing with the criticisms against prototype theory is embodied in exemplar theory (Brooks, 1978; Estes, 1986a, 1994, Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). Exemplar theory is based on two key assumptions:
1. **Representation**: a category is represented in memory as the set of representations of all category exemplars that have yet been encountered.4

2. **Access**: categorization decisions are based on similarity comparisons between the stimulus and the memory representation of *every* exemplar of each relevant category.

Two aspects of these assumptions are especially controversial. First, the assumption that categorization depends exclusively on exemplar or episodic memory has recently been called into question. A series of neuropsychological studies has shown that amnesic patients, with impaired episodic memory, can perform normally on a number of different category learning tasks (Knowlton, Ramus, & Squire, 1992; Knowlton & Squire, 1993; Kolodny, 1994). Second, the assumption that the similarity computations include *every* exemplar of the relevant categories is often regarded as intuitively unreasonable. For example, Myung (1994) argued that “it is hard to imagine that a 70 year-old fisherman would remember ever instance of fish that he has seen when attempting to categorize an object as a fish” (p. 348). Even if the exemplar representations are not consciously retrieved, a massive amount of activation is assumed by exemplar theory. One possibility that would retain the flavor of the exemplar approach is to assume that some random sample of exemplars are drawn from memory and similarity is only computed between the stimulus and this reduced set. However, one advantage of exemplar theory over prototype theory is that it can account for the observed sensitivity of people to the correlational structure of a category (Medin et al., 1982; Medin & Schaffer, 1978). It displays this sensitivity because the entire category is sampled on every trial. If only a subset of the category exemplars are sampled, the resulting model must necessarily be less sensitive to correlational structure. Ennis and Ashby (1993) showed that if only a single random sample is drawn, then exemplar models are relatively insensitive to correlational structure. To date, no one has investigated the question of how small the random sample can be, before adequate sensitivity to correlation is lost.

Many different exemplar models have been proposed. Figure 3 presents the hierarchical relations among some of these, as well as the relations among other types of categorization models. Models that are higher up in the tree are more general, whereas those below are special cases.

Perhaps the most prominent of the early exemplar models are the context model (Medin & Schaffer, 1978), the array-similarity model (Estes, 1994; also called the basic exemplar-memory model, Estes, 1986a), and the generalized context model (GCM; Nosofsky, 1985, 1986). In addition to the two

---

4 This assumption does not preclude the possibility of decay in the information with time or that only partial exemplar information is stored.
assumptions listed earlier, these models all assume the probabilistic response selection rule of the similarity-choice model—that is, Eq. (1)—although the context and array-similarity models allow no response bias. These latter two models also assume that all psychological dimensions are binary valued, and that the similarity between a stimulus and a stored exemplar is multiplicative across dimensions.\(^5\) In other words, if there are \(m\) dimensions, the similarity between stimulus \(i\) and stored exemplar \(j\) equals

\[
S_{ij} = \prod_{k=1}^{m} s_k(i, j),
\]

where \(s_k(i, j)\) is the similarity between the dimension \(k\) values of stimulus \(i\) and exemplar \(j\). From Assumption (2), all exemplar models assume the overall similarity of stimulus \(i\) to category \(j\) equals

\[
S_j = \sum_{j \in J} S_{ij}.
\]

\(^5\) It is possible to define an exemplar model in which similarity is additive. However, as shown by Nosofsky (1992b), an additive similarity exemplar model is equivalent to an additive similarity prototype model.
In the context model, the component similarity function is defined as

$$s_b(i, j) = \begin{cases} 1, & \text{if } i = j \\ q_b, & \text{if } i \neq j \end{cases}$$

The array-similarity model uses the same definition, except the similarity parameter $q_b$ is assumed to be the same on every dimension (i.e., $q = q_1 = q_2 = \ldots = q_m$).

The generalized context model (GCM) assumes continuous-valued dimensions. Similarity is defined flexibly and, as a result, only some versions of the GCM assume similarity is multiplicative across dimensions (Nosofsky, 1984). In all versions, however, the component similarity function, that is, $s_b(i, j)$, decreases symmetrically with distance from the stimulus representation. The model has two types of parameters that can stretch or shrink the psychological space. An overall discriminability parameter, $\epsilon$, expands or contracts the space uniformly in all directions. The attention weight parameters, $w_b$, selectively stretch or shrink dimension $k$ only. As a subject gains experience with a stimulus set, individual stimuli begin to look more distinct and, as a result, the similarity between a fixed pair of stimuli should decrease with experience. The GCM models this phenomenon by increasing the overall discriminability parameter $\epsilon$ with the subject’s level of experience (e.g., Nosofsky et al., 1992). Increasing a specific attention weight, say, $w_b$, stretches dimension $k$ relative to the other psychological dimensions. This selective stretching acts to decrease the dimension $k$ component similarities. The idea is that with more attention allocated to dimension $k$, the subject is better able to discriminate between stimuli that have different values on that dimension.

Although the context model has no attention weight or overall discriminability parameters, it is able to mimic the effects of these parameters by changing the magnitudes of the component similarity parameters (i.e., the $q_b$). Decreasing all $q_b$ by the same proportion is equivalent to increasing overall discriminability, and decreasing a single $q_b$ is equivalent to increasing a single attention weight. Under this interpretation, the assumption of the array-similarity model that all $q_b$ are equal is equivalent to assuming that equal amounts of attention are allocated to all stimulus dimensions.

The context and generalized context models have been used to account for asymptotic categorization performance from tasks in which the categories (1) were linearly or nonlinearly separable (Medin & Schwanenflugel, 1981; Nosofsky, 1986, 1987, 1989), (2) differed in base rate (Medin & Edelson, 1988), (3) contained correlated or uncorrelated features (Medin et al.,

---

6 Hintzman’s (1986) MINERVA2 is similar to the context and array-similarity models. All three make identical representation assumptions, although MINERVA2 does not assume a multiplicative similarity rule.
1982), (4) could be distinguished using a simple verbal rule (or a conjunction of simple rules; Nosofsky, Clark, & Shin, 1989), and (5) contained differing exemplar frequencies (Nosofsky, 1988a). The array-similarity model was developed primarily to predict category learning. The model has been applied to learning data in which the categories (1) were defined by independent or correlated features (Estes, 1986b) and (2) differed in base rate (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989). Recently, Estes (1994; see also Nosofsky et al., 1992) elaborated the model to predict a wider range of category learning phenomena.

In experiments where the stimuli are constructed from continuous-valued dimensions, unique parameter estimation problems are encountered. For example, in the GCM, the coordinates in psychological space of every category exemplar are free parameters (as well as the attention weights, overall discriminability, and response biases). Estimation of all these parameters requires many more degrees of freedom than are found in a typical categorization experiment. Nosofsky (1986) discovered an interesting solution to this problem. In a typical application, the coordinates of the stimuli in the psychological space are first estimated from data collected in a similarity judgment or stimulus identification task. Next, a recognition memory, typicality rating, or categorization task is run with the same stimuli. The GCM is then fit to this new data under the assumption that the stimulus coordinates are the same in the two experiments (see Nosofsky, 1992a for a review).

Ashby and Alfonso-Reese (1995) showed that the context model, the array-similarity model, and the GCM are all mathematically equivalent to a process in which the subject estimates the category likelihoods with a powerful nonparametric probability density estimator that is commonly used by statisticians (i.e., a Parzen, 1962, kernel estimator). This means that with a large enough sample size, these models can recover any category distribution, no matter how complex. The only requirements are that the subject does not completely ignore any stimulus dimensions and that overall discriminability slowly increases with sample size (as in Nosofsky et al., 1992). Thus, in most applications, the only suboptimality in these exemplar models that cannot be overcome with training is that they assume a probabilistic decision rule instead of the deterministic rule of the optimal classifier. In other words, the assumptions of exemplar theory, as embodied in the context model, the array-similarity model, or the GCM, are equivalent to assuming that the subject estimates all the relevant category distributions with an extremely powerful probability density estimator. The estimator is so powerful that it is bound to succeed, so these exemplar models predict that subjects should eventually learn any categorization problem, no matter how complex.

Recently, there has been a surge of interest in developing and testing models of category learning. The context, array-similarity, and generalized
context models provide adequate descriptions of asymptotic categorization performance, but these models are severely limited in their ability to account for the dynamics of category learning. The models have two main weaknesses. First, they all assume that the memory strength of an item presented early in the learning sequence remains unchanged throughout the course of the experiment. Thus, the influence of early items on performance during the last few trials is just as strong as the influence of items presented late in the learning sequence. Yet recency effects are well established in the memory literature—that is, a recently presented item will have a larger effect on performance than an item presented early in learning. To account for recency effects in category learning, exemplar theorists proposed that the memory strength of an exemplar decreases with the number of trials since it was last presented as a stimulus (Estes, 1993, 1994; Nosofsky et al., 1992).

A second problem with the context, array-similarity, and generalized context models is that they predict categorization response probabilities early in the learning sequence that are more extreme (i.e., closer to 0 and 1) than those observed in the empirical data. To see this, consider a categorization task with two categories, $A$ and $B$. If the first stimulus in the experiment is from category $A$, then the models predict that the probability of responding $A$ on the second trial is 1. The empirical data, on the other hand, suggest that early in the learning sequence, response probabilities are close to 0.5. To deal with this weakness, exemplar theorists postulated that subjects enter a category learning task with some information already present in the memory array that they will use for the representation of the contrasting categories. This background noise (Estes, 1993, 1994; Nosofsky et al., 1992) is assumed to be constant across categories and remains unchanged throughout the learning sequence. Early in the sequence, exemplar information is minimal and the background noise dominates, so categorization response probabilities are near 0.5. As more exemplars are experienced, the exemplar information in the memory array begins to dominate the background noise in the computation of the category response probabilities. Following Estes (1993, 1994), we will refer to context or array-similarity models that have been augmented with memory decay and background-noise parameters as the exemplar-similarity model (Ex-Sim; also called the sequence-sensitive context model by Nosofsky et al., 1992).

Another set of category learning models have been implemented in connectionist networks (e.g., Estes, 1993, 1994; Estes et al., 1989; Gluck & Bower, 1988; Gluck, Bower, & Hee, 1989; Hurwitz, 1990; Kruschke, 1992; Nosofsky et al., 1992). These models assume that a network of nodes and interconnections is formed during learning. The nodes are grouped into layers and information from lower layers feeds forward to the next higher layer. The input layer consists of nodes that correspond to individual
features, or collections of features (possibly even complete exemplar patterns). The output layer has a node associated with each of the contrasting categories. The amount of activation of a category (i.e., output) node is taken as the strength of association between the stimulus and that particular category. In most models, response selection is probabilistic. Whereas the exemplar-similarity model learns through a gradual accumulation of exemplar information, the connectionist models learn by modifying the weights between nodes as a function of error-driven feedback.

One of the earliest connectionist models of category learning was Gluck and Bower's (1988) adaptive network model. This is a feature-frequency model instantiated in a two-layer network. Gluck et al. (1989) proposed a configural-cue network model that generalizes the adaptive network model by including input layer nodes that correspond to single features, pairs of features, triples of features, and so on. Several exemplar-based connectionist models have also been developed. Estes (1993, 1994) proposed a two-layer connectionist model, called the similarity-network model (or Sim-Net), in which the input layer consists of exemplar nodes only. The hidden pattern unit model (HPU; Hurwitz, 1990, 1994) and ALCOVE (Kruschke, 1992) are three-layer networks in which the input layer consists of stimulus feature nodes, the second layer consists of exemplar nodes, and the output layer consists of category nodes.

Gluck and Bower (1988) reported data exhibiting a form of base-rate neglect that is predicted by the adaptive network models, but which has proved troublesome for the exemplar-similarity models. Subjects were presented with a list of medical symptoms and were asked to decide whether the hypothetical patient had one of two diseases. One of the diseases occurred in 75% of the hypothetical patients and the other disease occurred in 25% of the patients. After a training session, subjects were asked to estimate the probability that a patient exhibiting a particular symptom had one of the two diseases. On these trials, Gluck and Bower found that subjects neglected to make full use of the base-rate differences between the two diseases (see, also, Estes et al., 1989; Medin & Edelson, 1988; Nosofsky et al., 1992). This result is compatible with several different adaptive network models and incompatible with the exemplar-similarity model. For several years, it was thought that ALCOVE could account for the Gluck and Bower form of base-rate neglect (Kruschke, 1992; Nosofsky et al., 1992), but Lewandowsky (1995) showed that this prediction holds only under a narrow set of artificial circumstances. Thus, it remains a challenge for exemplar-based learning models to account for the results of Gluck and Bower (1988).

Although the exemplar-similarity model and the exemplar-based connectionist models have each had success, neither class has been found to be uniformly superior. In light of this fact, Estes (1994) attempted to identify experimental conditions that favor one family of models over the other by
comparing their predictions across a wide variety of experimental situations. Although a review of this extensive work is beyond the scope of this chapter, some of the experimental conditions examined by Estes (1994) include manipulations of category size, training procedure, category confusability, repetition and lag effects, and prototype learning. Estes (1986b; 1994; see also Estes & Maddox, 1995; Maddox & Estes, 1995) also extended the models to the domain of recognition memory.

Figure 3 shows Anderson's (1990, 1991) rational model as a special case of exemplar theory. This is not exactly correct because the rational model does not assume that the subject automatically computes the similarity between the stimulus and all exemplars stored in memory. When the coupling parameter of the rational model is zero, however, each exemplar forms its own cluster (Nosofsky, 1991a) and, under these conditions, the rational model satisfies our definition of an exemplar model. If, in addition, the similarity between the category labels is zero, and the subject bases his or her decision solely on the stored exemplar information, then the rational model is equivalent to the context model (Nosofsky, 1991a). Figure 3 also shows the prototype models to be special cases of the rational model. When the value of the coupling parameter in the rational model is one, and the similarity between the category labels is zero, the rational model reduces to a multiplicative similarity prototype model (Nosofsky, 1991a), such as Massaro's (1987) fuzzy logical model of perception.

E. Decision Bound Theory

The final theoretical perspective that we will discuss is called decision bound theory or sometimes general recognition theory (Ashby, 1992a; Ashby & Gott, 1988; Ashby & Lee, 1991, 1992; Ashby & Maddox, 1990, 1992, 1993; Ashby & Townsend, 1986; Maddox & Ashby, 1993). As described in the representation section, decision bound theory assumes the stimuli can be represented numerically but that there is trial-by-trial variability in the perceptual information associated with each stimulus, so the perceptual effects of a stimulus are most appropriately represented by a multivariate probability distribution (usually a multivariate normal distribution). During categorization, the subject is assumed to learn to assign responses to different regions of the perceptual space. When presented with a stimulus, the subject determines which region the perceptual effect is in and emits the associated response. The decision bound is the partition between competing response regions. Thus, decision bound theory assumes no exemplar information is needed to make a categorization response; only a response label is retrieved. Even so, the theory assumes that exemplar information is available. For example, in recall and typicality rating experiments, exemplar information must be accessed on every trial. Even in categorization tasks, the subject
might use exemplar information between trials to update the decision bound.

Different versions of decision bound theory can be specified depending on how the subject divides the perceptual space into response regions. The five versions that have been studied are (1) the independent decisions classifier (IDC), (2) the minimum distance classifier (MDC), (3) the optimal classifier (OC), (4) the general linear classifier (GLC), and (5) the general quadratic classifier (GQC). An example of each of these models is presented in Figure 4 for the special case in which the category exemplars vary on two

![Decision Bound Models](image-url)

**FIGURE 4** Decision bounds from five different decision bound models. (a) independent decisions classifier, (b) minimum distance classifier, (c) optimal decision bound model, (d) general linear classifier, and (e) general quadratic classifier.
perceptual dimensions. Figure 4 also assumes the category representations are bivariate normal distributions, but the models can all be applied to any category representation. The ellipses are contours of equal likelihood for the two categories. Every point on the same ellipse is an equal number of standard deviation units from the mean (i.e., the category prototype) and is equally likely to be selected if an exemplar is randomly sampled from the category. The exemplars of category A have about equal variability on perceptual dimensions x and y but the values on these two dimensions are positively correlated. In category B there is greater variability on dimension x and the x and y values are uncorrelated.

The independent decisions classifier (Figure 4a; Ashby & Gott, 1988; Shaw, 1982) assumes a separate decision is made about the presence or absence of each feature (e.g., the animal flies or does not fly) or about the level of each perceptual dimension (e.g., the stimulus is large or small). A categorization response is selected by examining the pattern of decisions across the different dimensions. For example, if category A was composed of tall, blue rectangles, then the subject would separately decide whether a stimulus rectangle was tall and whether it was blue by appealing to separate criteria on the height and hue dimensions, respectively. If the rectangle was judged to be both tall and blue, then it would be classified as a member of category A. Using a set of dimensional criteria is equivalent to defining a set of linear decision bounds, each of which is parallel to one of the coordinate axes (as in Figure 4a). The decision rule of the independent decisions classifier is similar to the rule postulated by classical theory. In both cases the subject is assumed to make a separate decision about the presence or absence of each stimulus dimension (or feature). Thus, classical theory is a special case of the independent decisions classifier (Ashby, 1992a).

Recently, Nosofsky, Palmeri, and McKinley (1994) proposed a model that assumes subjects use independent decisions bounds but then memorize responses to a few exemplars not accounted for by these bounds (i.e., exceptions to the independent decisions rule). Category learning follows a stochastic process, so different subjects may adopt different independent decisions bounds and may memorize different exceptions. The model was developed for binary-valued dimensions. It is unclear how it could be generalized to continuous-valued dimensions, since there is an abundance of continuous-valued data that is incompatible with virtually all independent decisions strategies, even those that allow the subject to memorize exceptions (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992).

The minimum distance classifier (Figure 4b) assumes the subject responds with the category that has the nearest centroid. An A response is given to every stimulus representation that is closer to the category A centroid, and a B response is given to every representation closer to the B centroid. The decision bound is the line that bisects and is orthogonal to the
line segment connecting the two category centroids. If the category centroid is interpreted as the prototype and if the similarity between the stimulus and a prototype decreases with distance, then the minimum distance classifier is an MDS prototype model.

The optimal decision bound model (Figure 4c) was proposed as a yardstick against which to compare the other models. It assumes the subject uses the optimal decision rule, that is, Eq. (4), and that the category likelihoods are estimated without error. Suboptimality occurs only because of perceptual and criterial noise.

The most promising decision bound models are the general linear classifier (Figure 4d) and the general quadratic classifier (Figure 4e), which assume that the decision bound is some line or quadratic curve, respectively. These models are based on the premise that the subject uses the optimal decision rule and estimates the category density functions (i.e., the stimulus likelihoods) using a parametric estimator that assumes the category distributions are normal. The assumption of normality is made either because experience with natural categories causes subjects to believe that most categories are normally distributed or because subjects estimate category means, variances, correlations, and base rates and infer a category distribution in the optimal fashion (see the representation section). If the estimated category structures are reasonably similar, the resulting decision bound will be linear (hence the general linear classifier), and if the estimated covariance structures are different, the bound will be quadratic (hence the general quadratic classifier). Specialized versions of the general linear classifier also have been developed to investigate the optimality of human performance when category base rates are unequal (Maddox, 1995).

One of the greatest strengths of decision bound theory is that it can be applied in a straightforward fashion to a wide variety of cognitive tasks. For example, different versions of the theory have been developed for application to speeded classification (Ashby & Maddox, 1994; Maddox & Ashby, 1996), identification and similarity judgment (Ashby & Lee, 1991; Ashby & Perrin, 1988), preference (Perrin, 1992), and same-different judgment (Thomas, 1994). In addition, decision bound theory provides a powerful framework within which to study and understand interactions during perceptual processing (Ashby & Maddox, 1994; Ashby & Townsend, 1986; Kadlec & Townsend, 1992). Although the theory allows for changes in the perceptual representation of stimuli across different tasks, it is assumed that the most important difference between, say, categorization and identification is that the two tasks require very different decision bounds. Ashby and Lee (1991, 1992) used this idea successfully to account for categorization data from the results of an identification task that used the same subjects and stimuli.
VI. EMPIRICAL COMPARISONS

This section reviews empirical data collected in traditional categorization tasks that test the validity of the categorization theories described in section V. We will also try to identify general properties of the stimuli, category structure, and training procedures that are most favorable to each theory. These properties are outlined in Table 3. Although it is usually the case that conditions opposite to those described in Table 3 are problematic for the various theories, this is not always true. As a result, the conclusions expressed in Table 3 must be interpreted with care.

A. Classical Theory

Classical theory assumes categorization is a process of testing a stimulus for the set of necessary and sufficient features associated with each relevant category. As described earlier, classical models are a special case of the independent decisions classifier of decision bound theory (Ashby, 1992a, Ashby & Gott, 1988). Ashby and Gott (1988, Experiments 1 and 3) and Ashby and Maddox (1990, Experiments 3 and 4; see, also Ashby & Maddox, 1992) tested the hypothesis that subjects always use independent decisions classification by designing tasks in which another strategy, such as minimum distance classification, yielded higher accuracy than the independent decisions classifier. The results convincingly showed that subjects were

<table>
<thead>
<tr>
<th>Theory</th>
<th>Experimental conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>Stimuli constructed from a few separable dimensions</td>
</tr>
<tr>
<td></td>
<td>Inexperienced subjects</td>
</tr>
<tr>
<td></td>
<td>Optimal rule is independent decisions</td>
</tr>
<tr>
<td></td>
<td>Taxonomic or logically defined categories</td>
</tr>
<tr>
<td>Prototype</td>
<td>Stimuli constructed from many integral dimensions</td>
</tr>
<tr>
<td></td>
<td>Inexperienced subjects</td>
</tr>
<tr>
<td></td>
<td>Optimal rule is complex or minimum distance</td>
</tr>
<tr>
<td></td>
<td>More than two categories</td>
</tr>
<tr>
<td>Exemplar</td>
<td>Experienced subjects</td>
</tr>
<tr>
<td></td>
<td>Optimal rule is simple</td>
</tr>
<tr>
<td></td>
<td>Few category exemplars</td>
</tr>
<tr>
<td>Decision Bound</td>
<td>Optimal rule is linear or quadratic</td>
</tr>
</tbody>
</table>
not constrained to use the independent decisions classifier. Instead, the best first approximation to the data was the optimal decision bound model.

Although classical theory is easily rejected, it has a certain intuitive attractiveness. For many tasks it seems the correct theory (e.g., categorizing squares versus pentagons). What properties do tasks that seem to favor classical theory possess? First, the task should use stimuli constructed from a few perceptually separable components. A pair of components are perceptually separable if the perceptual effect of one component is unaffected by the level of the other, and they are perceptually integral if the perceptual effect of one component is affected by the level of the other (e.g., Ashby & Maddox, 1994; Ashby & Townsend, 1986; Garner, 1974; Maddox, 1992). Classical theory predicts that a subject's decision about a particular component is unaffected by the level of other components. Thus, the independent decisions postulated by classical theory is a natural decision strategy when the stimulus dimensions are perceptually separable. Ashby and Maddox (1990) showed that experienced subjects are not constrained to use an independent decisions strategy, even when the stimulus dimensions are separable. Independent decisions is rarely optimal, and as subjects gain experience in a task, their performance naturally improves. Thus, a second experimental prerequisite is that the subjects are inexperienced (or unmotivated). One way to prevent a subject from gaining the kind of detailed category representation that comes from experience is to withhold feedback on each trial as to the correct response. Indeed, in unsupervised categorization tasks, subjects almost always use simple dimensional rules of the type assumed by classical theory (Ahn & Medin, 1992; Imai & Garner, 1965; Medin, Wattenmaker, & Hampson, 1987; Wattenmaker, 1992).

Finally, there are a few rare cases in which the independent decisions classifier is nearly optimal. For example, Ashby (1992a, Fig. 16.5, p. 473) proposed a task with normally distributed categories in which independent decisions is optimal. In such cases, we expect the data of experienced, motivated subjects to conform reasonably well to the predictions of the independent decisions classifier (and hence, to classical theory). In most experiments, of course, no effort is made to select categories for which independent decisions is optimal. If categories are selected without regard to this property, then chances are poor that independent decisions will be optimal. The best chances, however, although still poor, occur with categories that are taxonomic (e.g., as are many in the animal kingdom) or logically defined (e.g., as is the category "square"), because such categories are frequently defined by a list of characteristic features.
B. Prototype Theory

Prototype theory assumes the strength of association between the stimulus and a category equals the strength of association between the stimulus and the category prototype. All other exemplars or characteristics of the category are assumed to be irrelevant to the categorization process. Dozens of studies have shown convincingly that subjects are not constrained to use only the category prototypes (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Maddox & Ashby, 1993; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Nosofsky, 1987, 1992a; Shin & Nosofsky, 1992). Some of these studies compared prototype models with exemplar models, others compared prototype models with decision bound models. In every case, the prototype models were rejected in favor of either an exemplar or decision bound model.

These studies show that prototype theory does not provide a complete description of human categorization performance, but they do not rule out the possibility that the prototype has some special status within the category representation. For example, the prototype is sometimes classified more accurately (e.g., Homa & Culicé, 1984; Homa et al., 1981) and seems less susceptible to memory loss with delay of transfer tests than other category exemplars (e.g., Goldman & Homa, 1977; Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Homa & Vosburgh, 1976). In addition, as described in sections V.A and V.B, large prototype effects are almost always found in recall and typicality rating experiments.

Two recent categorization studies, however, found few, if any, special advantages for the prototype (Ashby et al., 1994; Shin & Nosofsky, 1992). Ashby et al. (1994) examined categorization response time (RT) in five separate experiments that used three different kinds of stimuli. No prototype effects on RT were found in any experiment. Hypotheses that assumed RT was determined by absolute or by relative distance to the two prototypes were both rejected. The prototypes did not elicit the fastest responses and even among stimuli that were just as discriminative as the prototypes with respect to the categorization judgment, the prototypes showed no RT advantage. The best predictor of the data was an assumption that RT decreased with distance from the decision bound. Thus, the fastest responses were to the most discriminative stimuli (which are furthest from the bound).

Shin and Nosofsky (1992) conducted a series of experiments using random dot stimuli that manipulated experimental factors such as category size, time between training and test, and within-category exemplar frequency. Each of these manipulations have been found to affect categorization accuracy for the prototype (e.g., Goldman & Homa, 1977; Homa & Chambless, 1975; Homa & Culicé, 1984; Homa, Dunbar, & Nohre, 1991; Homa...
et al., 1973, 1981; Homa & Vosburgh, 1976). Shin and Nosofsky (1992) replicated some prototype effects (e.g., increases in accuracy to the prototype with category size) but not others (e.g., differential forgetting for the prototype). In each experiment, Shin and Nosofsky tested a combined exemplar-prototype model, which contained a mixture parameter that determined the separate contribution of exemplar and prototype submodels to the predicted response probability. The mixture model provided a significant improvement in fit over a pure exemplar model in only one case, suggesting little, if any, contribution of the prototype abstraction process.

How can the results of Ashby et al. (1994) and Shin and Nosofsky (1992) be explained in light of the large empirical literature showing prototype effects? First, the failure to find prototype effects in categorization tasks is not necessarily damaging to prototype theory. Ashby and Maddox (1994) showed that in a categorization task with two categories, the most popular versions of prototype theory predict no prototype effects. Specifically, in most prototype models the predicted probability of responding $A$ on trials when stimulus $i$ is presented increases with $S_{ia}/S_{ib}$, whereas the response time decreases. Ashby and Maddox (1994) showed that this ratio increases with the distance from stimulus $i$ to the minimum distance bound. Because the prototype is usually not the furthest stimulus from the decision bound in two-category tasks, prototype theory predicts that the highest accuracy and the fastest responding will not be to the category prototypes, but to the stimuli that are furthest from the minimum distance bound.

In a task with a single category, in which the subject's task is to decide whether the stimulus is or is not a member of that category, the category prototype is often the furthest exemplar from the decision bound. Thus, prototype enhancement effects found in such tasks may not be due to an over representation of the prototype but instead to the coincidental placement of the prototype with respect to the subject's decision bound. This hypothesis was strongly supported by Ashby et al. (1994). Before deciding that prototype enhancement has occurred in a categorization task, it is vital to rule out the possibility that the superior performance to the prototype was not simply an artifact of the structure of the contrasting categories. Most studies reporting prototype enhancement in categorization tasks have not included analyses of this type, so the magnitude, and perhaps even the existence, of prototype effects in categorization is largely unknown.

In addition to category structure, other experimental conditions may make prototype effects more or less likely. Table 3 lists several properties that might bias an experiment in favor of prototype theory and thus might make prototype effects more likely. The minimum distance classifier, which uses the decision rule of prototype theory, generally requires the subject to integrate information across dimensions (e.g., Ashby & Gott, 1988). Although subjects can integrate information when the stimulus di-
dimensions are either integral or separable (Ashby & Maddox, 1990), integration should be easier when the stimulus dimensions are perceptually integral. Thus, it seems plausible that prototype theory might perform better when the category exemplars vary along perceptually integral dimensions (however, see Nosofsky, 1987).

Minimum distance classification is rarely optimal. Thus, as with independent decisions classification, any experimental conditions that facilitate optimal responding will tend to disconfirm prototype theory. Optimal responding is most likely when subjects are experienced (and motivated), the stimuli are simple, and the optimal rule is simple. Therefore, subjects should be most likely to use a suboptimal rule such as minimum distance classification if they are inexperienced, if the stimuli vary on many dimensions, if the optimal rule is complex, and if there are more than two categories (thus further complicating the optimal rule).

Much of the empirical support for prototype theory is from experiments that used random dot patterns as stimuli (e.g., Goldman & Homa, 1977; Homa et al., 1973, 1981; Homa & Culice, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968, 1970). These stimuli vary along many dimensions that are most likely integral. In addition, most of these experiments tested subjects for only a single experimental session, and thus the subjects were relatively inexperienced. Finally, a number of these experiments used more than two categories (e.g., Homa et al., 1973; Homa & Culice, 1984), so the experimental conditions were favorable for prototype theory.

Exemplar theorists have also questioned whether the existence of prototype effects necessarily implies that the category representation is dominated by the prototype. They argue that prototype effects are the natural consequence of exemplar-based processes of the kind hypothesized by exemplar theory. For example, Shin and Nosofsky (1992) found that the small prototype effects found in their categorization task could be predicted by an exemplar model. In addition, a number of investigators have shown that many of the prototype effects found in typicality rating, recognition, and recall tasks are qualitatively consistent with predictions from exemplar models (e.g., Bussemeyer, Dewey, & Medin, 1984; Hintzman, 1986; Hintzman & Ludlam, 1980; Nosofsky, 1988b).

Another hypothesis that explains the failure of Ashby et al. (1994) and Shin and Nosofsky (1992) to find prototype effects is that prototype effects are small or nonexistent in the majority of categorization tasks, but are robust in other types of cognitive and perceptual tasks. For example, the prevalence of prototype effects (or graded structure) in typicality rating tasks is uncontested. Graded structure has been found in a wide range of category types, for example, in taxonomic categories such as fruit (Rips et al., 1973; Rosch, 1973, 1975, 1978; Rosch & Mervis, 1975; Smith, Shoben, & Rips, 1974), logical categories such as odd number (Armstrong, Gietman,
Experiment 2 of McKinley and Nosofsky (1995) because of the many category exemplars and the complex optimal decision rule.

The general quadratic classifier of decision bound theory will give a good account of any data in which the subject uses a linear or quadratic decision bound (because the general linear classifier is a special case). Thus, if the optimal rule is linear or quadratic and the subjects are experienced, the general quadratic classifier should always perform at least as well as the best exemplar models. If the subjects respond optimally in a task where the optimal rule is more complex than any quadratic equation, then the general quadratic classifier will fit poorly (as in the McKinley & Nosofsky, 1995, experiments).

VII. FUTURE DIRECTIONS

Much is now known about the empirical validity of the various categorization theories and about their theoretical relations. As a result, the direction of research on human categorization is likely to change dramatically during the next decade. For example, advances in the neurosciences may make it possible to test directly some of the fundamental assumptions of the theories. In particular, through the use of various neuroimaging techniques and the study of selective brain-damaged populations, it may be possible to test whether subjects access exemplar or episodic memories during categorization, as assumed by exemplar theory, or whether they access some abstracted representation (e.g., a semantic or procedural memory), as assumed by decision bound theory. The early results seem problematic for exemplar theory, but the issue is far from resolved (Kolodny, 1994; Knowlton et al., 1992).

A second major distinction between exemplar theory and, say, the general quadratic classifier of decision bound theory is which a priori assumptions about category structure the subject brings to the categorization task. Exemplar theory assumes the subject makes almost no assumptions. When learning about a new category, exemplar theory assumes the subject ignores all past experience with categories. The general quadratic classifier assumes the subject brings to a new categorization task the expectation that each category has some multivariate normal distribution. Therefore, an extremely important research question is whether subjects make a priori assumptions about category structure, and if they do, exactly what assumptions they make.

A third important research question concerns optimality. Exemplar theory essentially assumes optimality for all categorization tasks, at least if the subjects have enough experience and motivation. The general quadratic classifier assumes optimality is possible only in some tasks (i.e., those in which the optimal bound is linear or quadratic, or possibly piecewise linear
or piecewise quadratic). Are there categorization problems that humans cannot learn? If so, how can these problems be characterized?

A fourth important direction of future research should be to explicate the role of memory in the categorization process. Because prototypicality effects seem to depend on the delay between training and test and on category size, it seems likely that memory processes play a key role in the development of the category prototype. A likely candidate is consolidation of the category representation. If so, then it is important to ask whether other observable effects of consolidation exist.

A fifth goal of future research should be to develop process models of the categorization task (i.e., algorithmic level models). The major theories reviewed in this chapter all have multiple process interpretations. A test between the various interpretations requires fitting the microstructure of the data. Simply fitting the overall response proportions is insufficient. For example, a process model should be able to predict trial-by-trial learning data and also categorization response times.

Finally, we believe that theories of human categorization would benefit greatly from the study of categorization in animals, and even in simple organisms. The first living creatures to evolve had to be able to categorize chemicals they encountered as nutritive or aversive. Thus, there was a tremendous evolutionary pressure that favored organisms adept at categorization. Because it is now believed that all organisms on Earth evolved from the same ancestors (e.g., Darnell, Lodish, & Baltimore, 1990), it makes sense that the categorization strategies used by all animals evolved from a common ancestral strategy. If so, then it is plausible that the fundamental nature of categorization is the same for all animals and that the main difference across the phylogenetic scale is in the degree to which this basic strategy has been elaborated (Ashby & Lee, 1993).

Acknowledgments

Preparation of this chapter was supported in part by National Science Foundation Grant DBS92-09411 to F. Gregory Ashby and by a Faculty-Grant-in-Aid from Arizona State University to W. Todd Maddox. We thank Kyunghee Koh and Leola Alfonso-Reese for their helpful comments on an earlier draft of this chapter.

References


Nosofsky, R. M. (1987). Attention and learning processes in the identification and categoriza-


& Gleitman, 1983), linguistic categories (see Lakoff, 1986 for a review), and many others (Barsalou, 1983, 1985). As discussed earlier, recognition memory (Omolunmoro, 1981), and recall (Mervis et al., 1976) also show marked prototype effects.

One thing that typicality rating, recognition memory, and recall tasks have in common is that they all require the subject to access exemplar information from memory. It is a matter of debate whether traditional categorization tasks require exemplar information (e.g., Ashby & Lee, 1991, 1992, 1993; Maddox & Ashby, 1993). Thus, one plausible hypothesis is that the prototype dominates the category representation, so any task requiring the subject to access the category representation will show prototype effects. Categorization experiments usually do not show prototype effects because categorization does not require information about individual exemplars. Why is the prototype overrepresented? One possibility is that the dominance of the prototype within the category representation is a consequence of consolidation processes. During periods of time in which the subject is gaining no new information about the category, the subject's memory for the category consolidates. The prototype begins to dominate the representation. According to this hypothesis, the prototype's prominence should increase with time because the consolidation process would have longer to operate. This prediction is consistent with the result that the few prototype effects that are found in categorization tasks tend to increase with the length of the delay between the training and testing conditions (e.g., Homa & Culicke, 1984). This result is important because it seems inconsistent with the hypothesis that prototype effects are the result of exemplar-based similarity computations.

C. Exemplar and Decision Bound Theory

In virtually every empirical comparison, exemplar models and decision bound models have both outperformed classical models and prototype models. There have been only a few attempts to compare the performance of exemplar and decision bound models, however. When the category distributions were bivariate normal and the data were analyzed separately for each subject, Maddox and Ashby (1993) found that the general linear and general quadratic decision bound models consistently outperformed Nosofsky's (1986) generalized context model. In cases where the optimal decision bound was linear, the advantage of the decision bound model was due entirely to response selection assumptions. A more general form of the generalized context model that used the deterministic Eq. (3) response selec-

7 Although nearly all categories show graded structure, Barsalou (1983, 1985, 1987) showed convincingly that graded structure is unstable and can be greatly influenced by context.
Although more empirical testing is needed, much is now known about how to design an experiment to give either exemplar or decision bound models the best opportunity to provide excellent fits to the resulting data. With respect to exemplars, the key theoretical result is that exemplar models (e.g., the model of Alfonso-Reese’s demonstration) are especially well fitted by the exemplar model, which is essentially an assumption that the subject is an extremely sophisticated statistician with perfect memory. Although exemplar models almost always predict that with enough training subjects will perform optimally, no matter how complex the task, they require either exemplar knowledge (e.g., the subject might be able to memorize the correct response to individual stimuli) or there are only a few exemplars in each category. This would allow nearly optimal responding, even in cases where the subject never learns the optimal rule. These conditions are summarized in Table 3.

Presumably, exemplar models performed poorly in