Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures

Stephen C. McKinley and Robert M. Nosofsky
Indiana University Bloomington

Experiments involving large-size, ill-defined categories were conducted to distinguish between the predictions of an exemplar model and linear and quadratic decision bound models. In conditions in which the optimal classification boundary was of a more complex form than the quadratic model, the exemplar model provided significantly better accounts of study participants' data than did the decision bound models, even in situations in which a linear bound would have yielded nearly optimal performance. The results suggest that participants are not predisposed or constrained to use linear or quadratic decision bounds for classifying multidimensional perceptual stimuli and that exemplar models may provide a parsimonious process-level account of the complex types of decision bounds used by experiment participants. The results also suggest some limitations on the complexity of the decision bounds that can be learned, in contrast to the predictions of the exemplar model.

A central property of most natural categories is that they have fuzzy, ill-defined structures. Objects in the real world display nearly infinite variability in their characteristics and, as a result, differ in the extent to which they are representative of alternative categories (Neisser, 1967; Rosch, 1973; Smith & Medin, 1981).

Because understanding how humans classify objects in the real world is of central importance in psychological research on perceptual categorization, an enormous amount of work has been devoted to the study of behavior involving ill-defined categories. Beginning with the influential work of researchers such as Posner and Keele (1968), Reed (1972), and Medin and Schaffer (1978), numerous theories and formal models of ill-defined classification have been advanced and tested over the past few decades. Most formal models of ill-defined classification have focused on similarity relations between objects and some underlying category representation. Possible choices for the underlying category representation in these tasks have included category prototypes (Homa & Chambliss, 1975; Posner & Keele, 1968; Reed, 1972), individual exemplars (Brooks, 1978; Hintzman, 1986; Medin & Schaffer, 1978), and feature-set tabulations (Hayes-Roth & Hayes-Roth, 1977; Neumann, 1974).

The assumption, referred to as exemplar theory, that categories are represented in memory by individual exemplars has generally been the most successful of these possible choices at predicting phenomena in the categorization literature. Specific models based on exemplar theory have enjoyed a qualitative advantage over prototype models (Medin & Schaffer, 1978; Nosofsky, 1987; Shin & Nosofsky, 1992) and independent cue models (Medin, Altom, Edelson, & Freko, 1982; Nosofsky, 1987; Nosofsky, Kruschke, & McKinley, 1992) across a variety of experimental tasks. In addition, formal models, most notably the generalized context model (GCM), have done an excellent job of predicting the details of study participant behavior across a wide variety of paradigms and choices of stimuli (Nosofsky, 1987, 1988, 1989; Nosofsky et al., 1992).

Recently, a new and contrasting theory of classification behavior, known as decision bound theory, has been proposed as an alternative to the exemplar approach. This theory, developed by Ashby and associates (Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992), is based on the general recognition theory of Ashby and Townsend (1986). A primary assumption underlying decision bound theory is that perception is an inherently noisy process in that repeated presentations of the same stimulus do not always give rise to the same perceptual effect in one's cognitive system. When applied to categorization, decision bound theory assumes that through experience, the categorizer learns to partition perceptual space into regions and associates a particular category label with each region. The boundaries between these perceptual regions are known as decision bounds. Thus, when faced with a classification decision, the categorizer determines in which region a particular percept lies and emits the associated category label in a deterministic fashion. Because a decision boundary can take essentially any form, constraints are needed if the theory is to generate testable predictions and be falsifiable. In recent work, Ashby and his associates have tested models...
in which the decision boundaries are optimal, linear, or quadratic in form (Ashby & Lee, 1991; Ashby & Maddox, 1990, 1992; Maddox & Ashby, 1993).

The central goal of the present research is to develop experimental contrasts to distinguish between the predictions of the exemplar and decision boundary models. One approach to developing such contrasts, beyond comparing the models’ ability to predict the details of classification performance, might involve the collection of converging forms of evidence bearing on the nature of the category representation. For example, in previous work, Nosofsky (1988, 1991; Shin & Nosofsky, 1992) demonstrated that the exemplar-based GCM provided excellent quantitative accounts not only of categorization but also of patterns of old–new recognition data observed at the completion of learning. It is unclear how a pure decision bound model would handle problems such as characterizing the categorization–recognition relation. Nevertheless, a decision bound theorist could argue that although decision boundaries are used for making classification judgments, exemplars are also stored as part of the learning process, and these exemplars are used for making old–new recognition judgments. Although such a dual model lacks parsimony, it cannot be ruled out. Therefore, the focus of the present work is to contrast the exemplar and decision boundary models solely on their ability to predict the details of classification performance.

Despite the fact that the exemplar and decision bound approaches differ greatly in their theoretical motivations and assumptions, specific models based on the theories have proven to be somewhat difficult to distinguish empirically. In one series of comparisons, Maddox and Ashby (1993) tested conditions in which the boundary between two categories that maximized classification accuracy was linear. In these conditions, a linear decision bound model and a version of the GCM with a deterministic response rule provided virtually equivalent accounts of study participant data. In another series of comparisons, Maddox and Ashby tested conditions in which the optimal classification boundary between two categories was quadratic in form. In these conditions, a decision bound model that postulated a quadratic decision bound, known as the general quadratic classifier (GQC), apparently outperformed the deterministic GCM.

The present article continues the line of investigation pursued by Maddox and Ashby (1993) in two ways. First, we demonstrate that the apparent advantage of the GQC reported by Maddox and Ashby disappears with the use of a more appropriate procedure for fitting the GCM. Second, we pursue model comparisons in a pair of experiments using category structures in which the optimal classification boundary is neither linear nor quadratic. This latter procedure is highly diagnostic for distinguishing between the predictions of the quadratic decision bound model and the exemplar model.

Viewed from the perspective of exemplar theory, the present experiments are of fundamental importance for two reasons. First, most previously successful applications of the exemplar-based GCM have occurred in experimental paradigms in which the number of exemplars per category is relatively small and in which there is a deterministic assignment of exemplars to categories. By contrast, in the present tests, we use an experimental paradigm developed by Ashby and his colleagues (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1992) in which the study participant experiences thousands of unique training exemplars and the assignment of exemplars to categories is probabilistic. Whereas the successful applications of the GCM in paradigms involving deterministic, small-size category structures may not be surprising to some investigators, the present experimental paradigm, with thousands of training exemplars assigned probabilistically to ill-defined structures, appears to provide a severe challenge to the generality of the approach. Second, whereas most of the previous tests of the GCM have involved fitting the model to data averaged over participants, the present tests involve the highly ambitious goal of fitting the model to individual participant classification data.

Viewed from the perspective of decision bound theory, the present experiments also address some issues of critical importance. Ashby and his associates (Ashby & Gott, 1988; Ashby & Maddox, 1992) have demonstrated convincingly that individuals are able to implement in a multidimensional perceptual space decision boundaries that are either linear or quadratic in form. First, one question that arises is whether subjects are predisposed to adopting decision boundaries at this level of complexity. In one of our experiments, we test a category structure in which a linear decision boundary yields levels of performance that are virtually as good as those produced by a much more complex optimal decision boundary. If the participants’ behavior is strongly determined by the complexity of the decision bound, we might expect them to use the simple linear boundary as a basis for classification. Second, our experiments explore whether the participants’ performance is constrained by the complexity of the boundary. In particular, will individuals adopt highly complex decision boundaries in situations in which they yield substantial improvements in classification performance, or are individuals’ abilities limited to adopting linear or quadratic decision boundaries?

In the next sections, we review the basic experimental paradigm in which the models are compared and then provide a brief review of the exemplar-based GCM and the quadratic decision bound model.

General Recognition Randomization Technique

The experimental paradigm in which all modeling comparisons take place is known as the general recognition randomization technique (GRRT). The GRRT paradigm was developed by Ashby and Gott (1988) as a means of observing the types of decision boundaries used by individuals when making classification decisions. In practice, the GRRT paradigm proceeds in the following way: First, the experimenter selects means and variance or covariance structures for a pair of category distributions (typically normal) along some predetermined stimulus dimensions. An example of the probability densities for two such bi-
variates normal category distributions is illustrated in Figure 1 (top). During a GRRT experimental session, category exemplars are randomly generated from the two distributions, and the study participant’s task is to classify the presented stimuli into one of the two categories. Corrective feedback is provided after every trial. Participants in GRRT paradigm experiments typically see between 1,200 and 2,000 category exemplars over a period of three to five experimental sessions. Note that because the distributions from which the stimuli are generated are continuous and overlapping, the category structure about which individuals must make decisions will be ill-defined and can, in theory, contain an infinite number of possible exemplars. As noted earlier, it is of fundamental interest whether exemplar models remain competitive under such conditions.

When describing the experiments that follow, it will be convenient to describe the category distributions used to generate exemplars by referring to what are known as equiprobability contours. An equiprobability contour is simply the set of points from a given distribution that are equally likely to occur. If the distributions are bivariate normal, the equiprobability contours will always be elliptical in shape. Examples of equiprobability contours from the two distributions specified in Figure 1 (top) are illustrated in Figure 1 (bottom) along with the optimal classification boundary. Equiprobability contours are a very convenient method of describing category structures in that they provide a vast amount of information concerning the category means and variance or covariance structures in a very compact visual form.

**Generalized Context Model**

The GCM (Nosofsky, 1984, 1986) is a generalization of the well-known context model of classification proposed by Medin and Schaffer (1978). The context model assumes that on a given trial, the classifier sums the similarity of a presented stimulus to the stored exemplars of each category and bases the classification decision on the relative magnitudes of these sums. Formally, in a two-category experiment, the model determines the probability that a pattern $i$ will be classified into Category A by summing the similarity of pattern $i$ to all previously stored exemplars of Category A and then dividing this sum by the summed similarity of pattern $i$ to all stored exemplars of both Categories A and B:

$$Pr(A|i) = \frac{B_A \cdot \Sigma s(i, a) + (1 - B_A) \cdot \Sigma s(i, b)}{B_A \cdot \Sigma s(i, a) + (1 - B_A) \cdot \Sigma s(i, b)},$$

(1)

where $s(i, a)$ represents the similarity of pattern $i$ to exemplar $a$ and $B_A$ ($0 \leq B_A \leq 1$) represents a response bias toward Category A.

A theoretical contribution made by Nosofsky (1984) was the interpretation of the similarity comparison process underlying Medin and Schaffer’s (1978) context model in terms of a multidimensional scaling approach. This contribution was formalized in the GCM and connected the categorization work of Medin and Schaffer with the influential similarity scaling work of Shepard (1964) and Garner (1974). In the GCM, the similarity between pattern $i$ and exemplar $a$ is interpreted as a decreasing function of distance in psychological space. Let $i_m$ and $a_m$ denote the psychological values on dimension $m$ of pattern $i$ and exemplar $a$. The distance in psychological space between these two objects is computed by using the weighted Minkowski power model formula

$$d(i, a) = \left(\sum |i_m - a_m|^r\right)^{1/r},$$

(2)

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

$$s(i, a) = \exp[-\kappa \cdot d(i, a)^p],$$

(3)

where $\kappa$ represents an overall sensitivity parameter. Common values of $p$ in Equation 3 are $p = 1$, which yields an exponential decay similarity gradient, and $p = 2$, which yields a Gaussian decay similarity gradient (Ennis, 1992; Nosofsky, 1985; Shepard, 1958, 1987). The predictions of the GCM are obtained by substituting the computed similarity values, $s(i, a)$, into the Equation 1 response rule. An important aspect of the GCM that we wish to emphasize here is that the similarity computations are influenced by a selective attention process that acts to differentially weight the psychological stimulus dimensions (Equation 2).
Deterministic Version of the GCM

In situations in which category exemplars are probabilistically assigned to categories (as in the GRRT paradigm), it is well-known that the GCM response rule (Equation 1) predicts probability matching behavior when similarities between all nonidentical patterns are zero (i.e., when \( \kappa = \infty \) in Equation 3) and predicts probability undermatching behavior when similarity relations are nonzero. The term probability matching behavior means simply that if a particular pattern \( i \) is presented as a member of Category A, say, 70% of the time, then the participant classifies pattern \( i \) into Category A 70% of the time.

Ashby and Gott (1988) provided convincing evidence that participants in the GRRT paradigm respond more deterministically (i.e., with probabilities closer to zero or unity) than is predicted by a probability matching rule. This evidence was further corroborated in direct tests that demonstrated failures in the quantitative predictions of the GCM (Maddox & Ashby, 1993). This shortcoming of the GCM led Nosofsky (1991) to propose a version of the model with a deterministic response rule. In this version, the basic spirit of the summed similarity comparison process is preserved, but the individual is assumed to respond deterministically with whichever category has the greater summed similarity. Noise in the system is placed at the level of the similarity computation process, but the response rule itself is deterministic.

Ashby and Maddox (1993) proposed a version of the GCM quite similar to that proposed by Nosofsky (1991), the main difference being that in Nosofsky’s model, the participant compares the summed similarities themselves, and in Ashby and Maddox’s model, the participant compares the natural logarithms of the summed similarities. The version developed by Ashby and Maddox is convenient in that it can be written as a direct generalization of the GCM (see Ashby & Maddox, 1993, for a proof). In this generalization, a parameter is added to the GCM response rule that allows the model to predict greater or lesser levels of response determinism than is predicted by the Equation 1 rule alone. Formally, the modified GCM response rule is given by

\[
Pr(A|i) = B_A \cdot [\sum M_s(i, a)]^\gamma / [B_A \cdot [\sum M_s(i, a)]^\gamma + (1 - B_A) \cdot [\sum M_s(i, b)]^\gamma],
\]

(4)

where \( \gamma \) reflects the amount of determinism in responding. Values of \( \gamma \) greater than 1 reflect greater levels of determinism than is predicted by Equation 1, and values of \( \gamma \) less than 1 reflect less determinism. We refer to the exemplar model using Ashby and Maddox’s modified response rule (Equation 4) as the GCM(D), where D stands for deterministic.

On the basis of the nature of the GRRT paradigm, we also explore an extended version of the GCM(D) with two additional free parameters. Typically, the psychological stimulus dimensions that are used in the GCM to calculate interpattern distances and similarities are independently derived from multidimensional scaling solutions (see Nosofsky, 1986, 1987, 1989). However, the huge number of individual stimuli that are generated in the GRRT paradigm precludes the possibility of deriving independent scaling solutions. Thus, in calculating similarities in the present paradigm, the stimulus coordinates used in the distance function (Equation 2) are simply the physical dimension values used to construct the stimuli. It is important to remember, however, that because the attention weights in the GCM are assumed to act on psychological, not physical, dimensions, it may be necessary to allow for reasonable transformations of the physical stimulus dimensions before distance and similarity calculations are performed. Maddox and Ashby (1993) tested a version of the GCM(D) using a transformation that allowed for a rotation of the physical dimensions about the origin of physical space prior to distance and similarity calculations, represented by a rotation parameter \( \theta \) (see Maddox & Ashby, 1993, for details). This transformation is well justified in the present context, and we decided to include the rotation parameter \( \theta \) in the elaborated GCM(D) model.

In addition, in a typical GRRT experiment, an individual views thousands of exemplars over a period of 5 days. It seems implausible that by the end of the experiment, an individual would actually be summing the similarity of a presented pattern to perfect memory traces of all previously seen exemplars. A straightforward method of addressing this concern is to postulate that more recently presented exemplars should receive greater weight in computing summed similarity. Recency effects are ubiquitous in the memory literature, and a version of the GCM incorporating the idea of recency effects provided a significant improvement over the standard GCM in predicting trial-by-trial learning data in previous research (Estes, 1994; Nosofsky et al., 1992). Thus, we hypothesized that the ability of the GCM(D) to predict the trial-by-trial learning data in the present study might be improved by the addition of a recency parameter. Following Nosofsky et al. (1992), the recency-sensitive version of the GCM(D) assumes that

\[
Pr(A|i) = B_A \cdot [\sum M_s(i, a)]^\gamma / [B_A \cdot [\sum M_s(i, a)]^\gamma + (1 - B_A) \cdot [\sum M_s(i, b)]^\gamma],
\]

(5)

where \( M_s \) is the memory strength associated with exemplar \( a \). Exemplar memory strength is assumed to decrease exponentially with lag of presentation:

\[
M_a = \exp(-T \cdot \text{lag}),
\]

(6)

where \( T \) is a freely varying trial rate parameter and lag is the number of intervening trials between the presentations of pattern \( i \) and exemplar \( a \). The full version of the GCM(D) model described above has six free parameters: a category bias parameter (\( B_A \)), an

---

1 In several of Maddox and Ashby’s (1993) experimental conditions, rectangular box stimuli were used in which the physical dimensions of length and width were varied. A good deal of empirical evidence exists that emergent psychological dimensions of shape and area may dominate individuals’ similarity judgments for such stimuli. The psychological dimensions of shape and area can be approximated by rotations of the length–width physical space.
attention weight parameter \((\alpha_1)\) that enters into the distance formula (Equation 2, \(\alpha_2 = 1 - \alpha_1\)), an overall sensitivity parameter \((\kappa)\) that enters into the similarity computation formula (Equation 3), a response determinism parameter \((\gamma)\) that enters into the response rule (Equation 4), a dimensional rotation parameter \((\theta)\), and a forgetting-rate parameter \((T)\).

**General Quadratic Classifier**

To motivate the GQC, Ashby (1992) and Ashby and Maddox (1992) suggested that the multivariate normal distribution provides a good approximation to the structure of numerous categories in the natural world. The multivariate normal consists of an unlimited number of unique exemplars, its dimensions are continuous valued, and it is unimodal and symmetric. Furthermore, the distributions of alternative multivariate normals are overlapping. These properties seem to hold true for numerous natural categories (see Ashby, 1992, and Fried & Holyoak, 1984, for more extensive discussions).

Suppose the assumption of multivariate normal categories in the natural world is roughly correct. Then it is reasonable to assume that as a result of their extensive experience with natural categories, study participants have learned to establish decision boundaries with the same form as those boundaries that are optimal for discriminating among the members of multivariate normal distributions. It is well-known that the optimal decision boundaries for discriminating the members of two multivariate normal distributions is either quadratic or linear in form. The optimal decision boundary is linear only in the special case in which the covariance matrices of the two distributions are identical (i.e., when the two distributions have the same variance on all dimensions and have the same covariances between dimensions). In all other cases the optimal decision boundary is quadratic.

On the basis of the considerations outlined above, Ashby (1992) proposed the GQC as a model of human categorization. It is critical to realize that the GQC does not assume that the individual necessarily adopts the optimal decision bound. Rather, it is assumed only that he or she adopts some quadratic boundary. For example, a participant may attempt to respond optimally but may underestimate the variability along one of the dimensions in a given category. In this case, he or she will use a quadratic bound, but not the one that is optimal for the classification task at hand. Maddox and Ashby (1993, p. 52) identified a variety of reasons why an individual may adopt a suboptimal quadratic decision boundary.

Formally, the GQC assumes that the participant constructs a quadratic decision boundary \(h(x)\) and uses the following decision rule to assign stimuli to categories:

\[
\begin{align*}
\text{if } h(x_{pl}) &< \delta; \text{ then respond A} \\
\text{if } h(x_{pl}) &= \delta; \text{ then guess} \\
\text{if } h(x_{pl}) &> \delta; \text{ then respond B,}
\end{align*}
\]

where \(\delta\) reflects a response bias, and the vector \(x_{pl}\) represents the individual’s perceptual representation of stimulus \(i\). The perceptual representation \((x_{pi})\) is assumed to be subject to noise in the processing system and is given by

\[
x_{pi} = x_i + e_p,
\]

where \(x_i\) is the mean percept of pattern \(i\) and \(e_p\) is a random vector that reflects perceptual noise. The perceptual noise vector \(e_p\) is assumed to be normally distributed with mean zero and covariance matrix given by \(\sigma_{e}^2 \cdot I\), where \(I\) is the appropriate identity matrix (Maddox & Ashby, 1993). The probability of responding A on trials when stimulus \(i\) is presented is

\[
Pr(A|i) = Pr(h(x_{pi}) < \delta + e_c),
\]

where \(e_c\) is a normally distributed random variable with mean zero and variance \(\sigma_{e}^2\) that reflects variability in the setting of the response bias criterion \(\delta\). Because \(e_c\) is symmetric about zero, Equation 9a can be rewritten as

\[
Pr(A|i) = Pr(h(x_{pi}) + e_c < \delta).
\]

The sum of \(h(x_{pi})\) and \(e_c\) is assumed to be normally distributed with mean value equal to that of \(h(x_{pi})\) and variance equal to the sum of the variances of \(h(x_{pi})\) and \(e_c\). Thus, the classification probabilities predicted by the GQC (Equation 9b) can be computed in a straightforward manner. With two perceptual dimensions, \(x_1\) and \(x_2\), the decision bound \(h(x_1, x_2)\) is

\[
h(x_1, x_2) = A_1 \cdot x_1^2 + A_2 \cdot x_2^2 + A_3 \cdot x_1 \cdot x_2 + A_4 \cdot x_1 + A_5 \cdot x_2 + A_6
\]

for some constants \(A_1 - A_6\). In the version of the GQC fitted throughout this article, there are seven freely varying parameters: five freely varying decision bound coefficients (one may be held fixed at some arbitrary value without loss of generality), a perceptual variability parameter \((\sigma_{x_i}^2)\), and a criterial variability parameter \((\sigma_{e}^2)\). Following Maddox and Ashby (1993), the response bias parameter \((\delta)\) was held fixed at zero. Allowing the parameter to vary has no effect on the model fits.

**Preliminary Model Comparisons**

As a preliminary step in comparing the GCM(D) and the GQC, we fitted the GCM(D) and the GQC to 23 sets of individual participant response data collected by Ashby and Maddox (1992) and 5 data sets collected by McKinley and Nosofsky (1994). Ashby and Maddox data were collected in a series of three standard GRRT experiments in which the optimal classification boundary was quadratic and in which no linear boundary provides even a rough

\footnote{Ashby and Maddox (1992) collected data from 24 individuals. However, the experimental data for 1 participant’s first session were inadvertently lost, so we were unable to fit the exemplar model to that individual’s incomplete data. The unpublished data collected in our laboratory are available from us on request. We are indebted to W. Todd Maddox for supplying the data sets from the Maddox and Ashby (1993) study.}
approximation to the optimal boundary. Equiprobability contours for the three category structures are illustrated in Figure 2. The three Ashby and Maddox experiments each used two types of stimuli: circles varying in radius with an embedded diameter varying in angle above horizontal and rectangular boxes varying in length and width. The McKinley and Nosofsky replication used only the Figure 2 (top) category structure and only the circular stimuli. For details of the experimental procedure, see Ashby and Maddox.\textsuperscript{3}

Maddox and Ashby (1993) previously reported fits of a five-parameter deterministic exemplar model (i.e., the GCM[D] described above without the forgetting parameter $T$) and the seven-parameter GQC to 24 data sets from the three Ashby and Maddox (1992) experiments. The data fitted in these comparisons were the classification responses observed during the final 300 trials of the experiment. In these comparisons it was shown that the GQC displayed a consistent advantage over the five-parameter deterministic GCM in that the GQC provided superior fits to 16 of 24 data sets. On the basis of these findings, it was concluded that the decision bound is a fundamental construct in predicting asymptotic categorization performance and that the GQC provides a viable alternative to exemplar models of classification.

Our reason for refitting the GCM(D) to these data sets is twofold. First, to predict the asymptotic classification data during the last 300 experimental trials using the exemplar model, one needs to sum the similarity of each of the final 300 patterns to the thousands of exemplars experienced during earlier experimental sessions. Obviously, this procedure has a high computational expense. As an approximation to fitting the exemplar model, therefore, Maddox and Ashby (1993) instead adopted the procedure of summing the similarity of the final 300 items to only those patterns presented during the final 410 trials. Unfortunately, this method of approximation means that in predicting classification for any given item, one is actually summing the similarity of that item to numerous patterns not yet experienced. Thus, although Maddox and Ashby provided evidence that their procedure yielded an adequate approximation, we felt it was important to fit the true version of the exemplar model. Therefore, we decided to refit all their data by adopting the expensive computational procedure of summing the similarity of each pattern to all previously seen exemplars.

Second, this procedure is necessary when fitting the recency-sensitive version of the GCM(D). According to this model, exemplars that have been presented in the more recent past receive greater weight in the summed similarity computations than those exemplars presented in the distant past. Therefore, to fit the recency-sensitive version of the model, by necessity one must sum the similarity of each pattern to all previously seen exemplars.

The fits of the models were computed using a hill-climbing algorithm that searched for parameter values that maximized the log likelihood of observing the set of responses made by a particular participant during the last 300 trials of the experiment (see Wickens, 1982, for an introduction to the method of maximum likelihood). The function used to determine the likelihood of observing a given set of responses $(r_1, \ldots, r_{300})$ is given by

$$L(r_1, \ldots, r_{300}) = \prod_{i=1}^{300} P(A_i | r_i) P(B | r_i)^{-1}$$

where $r_i = 1$ if response A was made to stimulus i and $r_i = 0$ if response B was made (Maddox & Ashby, 1993). To guard against settling into local minima, the search routine was initiated from three different locations in parameter space.

An obstacle in comparing the GCM(D) with the GQC is that the two models have different numbers of free parameters and the models are not hierarchically nested. Thus, straightforward comparisons based on maximum likelihood are not possible. However, statistical criteria have been developed that allow for comparisons of this type to be made (see Akaike, 1974; Stone, 1979). Following Maddox and Ashby (1993), the fits we report use a statistical criterion known as Akaike’s information criterion (AIC). The AIC is a modification of the method of maximum likelihood that penalizes models for extra free parameters. The AIC is defined as

$$\text{AIC}(M_i) = -2 \cdot L_i + 2 \cdot N_{\theta}$$

where $L_i$ is the maximum log likelihood of the data predicted by model $M_i$ and $N_{\theta}$ is the number of free parameters in model $M_i$. Smaller AIC values reflect a more accurate account of the data.

Table 1 reports the AIC fits of the GCM(D) and the GQC to the 28 data sets mentioned above.\textsuperscript{4} Following Maddox and Ashby (1993), the reported fits of the GCM(D) were computed by using a euclidean distance metric and an exponential decay similarity function. Of the 28 data sets, the GCM(D) produced the best AIC value for 17 cases, and the GQC produced the best AIC value in the other 11 cases. A restricted version of the GCM(D) in which the rotational parameter ($\theta$) and the forgetting-rate parameter ($T$) were held fixed at zero provided the best AIC fit to 15 of 28 data sets. As these results show, fitting the actual GCM(D) instead of the approximation version removes the advantage of the GQC reported in Maddox and Ashby. In fact, the four-parameter restricted GCM(D) is essentially equivalent to the GQC, with an average AIC value of 220.26 versus 220.04 for the GQC. The six-parameter GCM(D) displays a very slight advantage over the GQC with an average AIC value of 216.63, suggesting that the addition of the rotation parameter ($\theta$) and the forgetting-rate parameter ($T$) provides only a very modest improvement to the model’s performance. Perhaps the most reasonable conclusion that can be drawn is that the GCM(D) is a viable alternative to the GQC.

\textsuperscript{3} The experimental procedure used by McKinley and Nosofsky (1994) was identical to that used by Ashby and Maddox (1992), with the exception that the category prototypes and labels were not displayed prior to each experimental session. Also, the experimental stimuli consisted of circles with an embedded radial line as opposed to an embedded diameter.

\textsuperscript{4} Best fitting parameter values for these and all other fits in this article are available from us on request.
Table 1
Summary of Model Fits (Akaike’s Information Criterion) to the Ashby and Maddox (1992) and McKinley and Nosofsky (1994) Data Sets

<table>
<thead>
<tr>
<th>Experiment/participant</th>
<th>GQC(7)</th>
<th>GCM(D) 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN/1</td>
<td>327.82</td>
<td>283.62</td>
</tr>
<tr>
<td>MN/2</td>
<td>258.12</td>
<td>232.98</td>
</tr>
<tr>
<td>MN/3</td>
<td>173.70</td>
<td>231.28</td>
</tr>
<tr>
<td>MN/4</td>
<td>354.06</td>
<td>338.40</td>
</tr>
<tr>
<td>MN/5</td>
<td>271.90</td>
<td>270.42</td>
</tr>
<tr>
<td>1B/1</td>
<td>326.30</td>
<td>318.58</td>
</tr>
<tr>
<td>1B/2</td>
<td>199.90</td>
<td>185.70</td>
</tr>
<tr>
<td>1B/3</td>
<td>255.90</td>
<td>261.88</td>
</tr>
<tr>
<td>1B/4</td>
<td>273.30</td>
<td>259.48</td>
</tr>
<tr>
<td>1C/1</td>
<td>216.20</td>
<td>219.90</td>
</tr>
<tr>
<td>1C/2</td>
<td>217.90</td>
<td>197.68</td>
</tr>
<tr>
<td>1C/3</td>
<td>273.80</td>
<td>264.04</td>
</tr>
<tr>
<td>1C/4</td>
<td>290.70</td>
<td>296.40</td>
</tr>
<tr>
<td>2B/1</td>
<td>297.90</td>
<td>306.98</td>
</tr>
<tr>
<td>2B/2</td>
<td>259.60</td>
<td>258.28</td>
</tr>
<tr>
<td>2B/3</td>
<td>336.80</td>
<td>340.18</td>
</tr>
<tr>
<td>2C/1</td>
<td>212.20</td>
<td>212.90</td>
</tr>
<tr>
<td>2C/2</td>
<td>207.70</td>
<td>245.00</td>
</tr>
<tr>
<td>2C/3</td>
<td>215.20</td>
<td>211.90</td>
</tr>
<tr>
<td>2C/4</td>
<td>224.50</td>
<td>219.96</td>
</tr>
<tr>
<td>3B/1</td>
<td>112.50</td>
<td>95.52</td>
</tr>
<tr>
<td>3B/2</td>
<td>114.20</td>
<td>116.04</td>
</tr>
<tr>
<td>3B/3</td>
<td>133.90</td>
<td>110.98</td>
</tr>
<tr>
<td>3B/4</td>
<td>219.10</td>
<td>204.40</td>
</tr>
<tr>
<td>3C/1</td>
<td>128.80</td>
<td>137.08</td>
</tr>
<tr>
<td>3C/2</td>
<td>57.50</td>
<td>43.74</td>
</tr>
<tr>
<td>3C/3</td>
<td>122.30</td>
<td>127.44</td>
</tr>
<tr>
<td>3C/4</td>
<td>79.40</td>
<td>74.92</td>
</tr>
</tbody>
</table>

Note. Generalized context model (deterministic) [GCM(D)] fits computed with euclidean metric and exponential similarity decay. Fits of the general quadratic classifier (GQC) to the Ashby and Maddox (1992) data sets are from Maddox and Ashby (1993). The number in parentheses refers to the number of free parameters. MN = the McKinley & Nosofsky (1994) data sets; B = rectangular box stimuli; C = circular stimuli.


drawn from this analysis is that, overall, neither model displays a qualitative advantage over the other in these experiments.

We feel it is important to point out that, in fact, it would be very difficult to ever find convincing evidence in favor of the GCM(D) over the GQC in the category structures tested, even if the exemplar model comes close to being the "true" model. The reasoning behind this assertion is as follows: One can think of the GCM(D) as producing a "decision boundary" in much the same way that the GQC produces a decision boundary. That is, the GQC determines a boundary in space such that, for all points on one side of the boundary, the study participant responds with one category, and for all points on the other side, the participant responds with the alternative category. Likewise, the GCM(D) produces a set of points in space for which the summed similarities to both categories are identical. For all points on one side of this boundary, the participant responds with one category, and for all points on the other side, the participant responds with the alternative category. Ashby and Maddox (1993) referred to this set of points as the equivocality contour. It can be
shown that when the categories are normally distributed and the optimal classification boundary is quadratic, the asymptotic equivocality contour produced by the Gaussian or Euclidean GCM will always be quadratic in form. In addition, numerical investigations conducted in our laboratory have suggested that this contour is roughly quadratic in form under alternative choices of distance metric and similarity decay function. More important, the location in space of the quadratic equivocality contour produced by the GCM(D) in these structures is highly constrained by the set of training exemplars from which it is derived. Thus, it seems unlikely that the GCM(D) could ever be found to offer significant improvements in these structures over the GQC, which uses a freely varying quadratic decision boundary. In summary, the results of our preliminary model-fitting analyses, as well as the theoretical arguments just advanced, suggest that more diagnostic structures will be needed to significantly distinguish between these two modeling approaches.

**Experiment 1**

The main goal of Experiment 1 was to investigate more diagnostic structures for distinguishing between the GCM(D) and the GQC. As long as the optimal classification boundary remains quadratic or linear, both models appear to generate similar predictions of asymptotic classification probabilities. Thus, it is important to test the models in GRRT structures in which the optimal classification boundary is of a more complex form than quadratic.

To achieve this goal, Experiment 1 compares the abilities of the GQC and the GCM(D) in a GRRT paradigm experiment in which the categories are generated from mixtures of bivariate normal distributions. A bivariate normal mixture density \( f(x, y) \) is defined as

\[
f(x, y) = \sum p_i \cdot g_i(x, y),
\]

where \( g_i(x, y) \) is the joint density function for bivariate normal distribution \( i \) and \( p_i \) is known as the mixing proportion for distribution \( i \) (\( \sum p_i = 1 \)). The category structures used in Experiment 1 are illustrated in Figure 3. Each category was generated from a mixture composed of two bivariate normal densities with equal mixing proportions (i.e., \( p_1 = p_2 = 0.5 \) for each category). The actual means and variance or covariance structures for each normal mixing distribution \( g_i(x, y) \) in each of the two Experiment 1 conditions are given in Table 2. The Condition 2 category structure, which can be obtained by rotating the Condition 1 structure 45° about its center of mass, was designed to increase the generality of the results in the forthcoming modeling analyses.

The optimal classification boundaries for the Experiment 1 category structures are illustrated in Figure 3. Of critical importance to the present design is the fact that the optimal boundaries are nonquadratic. Although each boundary is locally quadratic over measurable regions of the stimulus space, the boundaries are globally complex combinations of quadratic curves. Thus, because the GCM(D) tends to predict decision boundaries with a form similar to the optimal bound regardless of the nature of the category distributions (e.g., Ashby & Maddox, 1993), the present structures should be highly diagnostic for distinguishing between the predictions of the GCM(D) and the GQC.

The present experimental design has some additional virtues. Visual inspection of the structures in Figure 3
Table 2
Category Distribution Parameter Values for Experiment 1

<table>
<thead>
<tr>
<th>Category</th>
<th>g(x, y)</th>
<th>μx</th>
<th>μy</th>
<th>σx</th>
<th>σy</th>
<th>ρxy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>120</td>
<td>80</td>
<td>35</td>
<td>35</td>
<td>−0.9</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>160</td>
<td>120</td>
<td>35</td>
<td>35</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>80</td>
<td>120</td>
<td>35</td>
<td>35</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>120</td>
<td>160</td>
<td>35</td>
<td>35</td>
<td>−0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>148</td>
<td>92</td>
<td>48.2</td>
<td>11.1</td>
<td>−0.9</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>148</td>
<td>148</td>
<td>11.1</td>
<td>48.2</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>92</td>
<td>92</td>
<td>11.1</td>
<td>48.2</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>92</td>
<td>148</td>
<td>11.1</td>
<td>48.2</td>
<td>−0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *g(x, y) = density function; μ = mean; σ = standard deviation; ρ = correlation. Subscript r refers to the radius of the circle; subscript a refers to the angle above horizontal. Following Maddox and Ashby (1993), we have adopted the convention of expressing angle values in “units,” where π/4 radians equals 120 units.

reveals that a linear boundary separates the categories quite well. Indeed, whereas a participant using the optimal pop-
ulation classification boundary illustrated in Figure 3 would correctly classify approximately 91% of the stimuli, the most accurate linear boundary in each condition correctly classifies 86% of the stimuli. Because the performance predicted by the most accurate linear boundaries is only 5% less than that predicted by the optimal likelihood bound-
aries, a reasonable hypothesis stemming from decision bound theory might be that participants would adopt a linear boundary and use it throughout the experiment. An impor-
tant theoretical issue in the upcoming modeling analyses, therefore, is to test if this hypothesis is supported by com-
paring the predictions of a linear decision bound model to those of the GQC and GCM(D).

Finally, we remark that mixture distributions such as those in Experiment 1 may provide a simplified model of the structure of many natural categories. If the structures of many natural categories are approximately multivariate nor-
mal, as argued by Ashby (1992), then it seems reasonable that many natural categories may also be composed of mixtures of normals. The reason is that many natural cat-
egories may be composed of subtypes, which would them-
selves be multivariate normal. To illustrate, consider the natural category deer. When deciding if an object in the environment is a member of the category deer, one is typically able to perform the task without knowledge of whether the object falls into the group male or female. Yet, it certainly seems reasonable that the distribution of (adult) deer differs among these subtypes. Thus, if one accepts that the multivariate normal is a good approximation to the structure of many natural categories, the assumption that many natural categories may be composed of mixtures of normals seems equally plausible.

Method

Participants. All participants were graduate students at Indiana University who were unfamiliar with the theoretical issues under investigation in this research. Graduate students were chosen to ensure a well-motivated participant population. A total of 5 stu-
dents participated in each condition, for which they were paid $5 per experimental session plus a bonus of 10¢ for each percentage point above 70% achieved during the last half of each session. The participants were paid in full on completion of the experiment.

Stimuli. The experimental stimuli were circles varying in ra-
dius with an embedded radial line varying in angle. The stimuli were computer generated on IBM 386 style PCs using BASIC and were displayed on a VGA monitor. The screen resolution was defined within the BASIC program to be 640 × 480 units. The units given in Table 2 correspond to the program-defined units above. These stimuli were chosen because there is extensive evi-
dence that their underlying psychological dimensions correspond closely to the physically manipulated dimensions. This close correspondence is necessary to achieve rigorous comparisons between the models in the present paradigm.

Procedure. Stimulus generation on each trial proceeded as outlined here. First, Category A or B was randomly selected with probability .5 each. Then, one of the two distributions from the selected category was randomly selected, and a stimulus was sampled from the selected distribution and presented on the screen. The participant’s task on each trial was to classify the presented stimulus into Category A or Category B. The display was response terminated. After the participant entered a response, the computer provided corrective feedback. Each experimental session consisted of four 200-trial blocks, and participants were allowed to take a short break between each block. At the end of each session, a message appeared on the screen informing the participants of their response accuracy over the last 400 trials of that day’s session; they were told before the task that even after extensive training an expert would be expected to make occasional errors. Each indi-
vidual participated in a total of five sessions over a contiguous work week, for a total of 4,000 experimental trials.

Results and Theoretical Analyses

Table 3 shows the percent correct achieved for each participant during each session. Performance levels appear to have asymptoted by the third or fourth session.

To provide some perspective on the types of asymptotic response patterns, Figures 4A and 4B show the responses made during the last 300 trials of the experiment by the

Table 3
Percentage Correct Over Each Experimental Session for Each Participant in Experiment 1

<table>
<thead>
<tr>
<th>Participant</th>
<th>Session</th>
<th>Condition 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>71</td>
<td>81</td>
<td>82</td>
<td>84</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>73</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>56</td>
<td>78</td>
<td>81</td>
<td>79</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>74</td>
<td>79</td>
<td>80</td>
<td>88</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>63</td>
<td>75</td>
<td>77</td>
<td>81</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Condition 2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>79</td>
<td>82</td>
<td>84</td>
<td>84</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>66</td>
<td>77</td>
<td>78</td>
<td>84</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>58</td>
<td>74</td>
<td>71</td>
<td>75</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>76</td>
<td>85</td>
<td>80</td>
<td>83</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>81</td>
<td>82</td>
<td>83</td>
<td>81</td>
<td>82</td>
<td></td>
</tr>
</tbody>
</table>
top-performing participants in each condition (Participant 4 in Condition 1 and Participant 4 in Condition 2). For convenience, the optimal classification boundaries from Figure 3 are once again illustrated. As shown, the participants’ responses tended to map fairly closely onto the actual category densities themselves and are fairly well separated by the optimal bound. It is also evident in Figure 4 that the response patterns exhibited by these participants do not appear to be well approximated by the assumption of a single linear or quadratic decision boundary.

Also shown in Figure 4 are the responses made by 2 other representative participants, Participant 2 in Condition 1 and Participant 2 in Condition 2. Although somewhat more noisy, the response patterns still display a sensitivity to the category distributions similar to that displayed by the top-performing participants. The only participant who deviated from this pattern was Participant 5 in Condition 2. This student’s performance asymptoted during the first experimental session, and visual inspection of the response pattern revealed that this individual may have adopted and used a
Table 4  
Model Fits (Akaike's Information Criterion) to Last 300 Trials of Final Experimental Session in Experiment 1

<table>
<thead>
<tr>
<th>Participant</th>
<th>GLIN(3)</th>
<th>GQC(7)</th>
<th>GCM(D) 6</th>
<th>QNET(10)</th>
<th>GDQC(12)</th>
<th>QUAR(15)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>170.52</td>
<td>149.86</td>
<td>118.24</td>
<td>130.44</td>
<td>147.08</td>
<td>123.66</td>
</tr>
<tr>
<td>2</td>
<td>201.28</td>
<td>203.98</td>
<td>188.54</td>
<td>195.14</td>
<td>197.42</td>
<td>200.12</td>
</tr>
<tr>
<td>3</td>
<td>268.62</td>
<td>265.92</td>
<td>253.12</td>
<td>246.88</td>
<td>250.44</td>
<td>244.18</td>
</tr>
<tr>
<td>4</td>
<td>210.74</td>
<td>192.84</td>
<td>114.90</td>
<td>164.46</td>
<td>122.80</td>
<td>119.04</td>
</tr>
<tr>
<td>5</td>
<td>217.28</td>
<td>214.90</td>
<td>181.92</td>
<td>198.78</td>
<td>188.10</td>
<td>200.70</td>
</tr>
<tr>
<td><strong>Condition 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>255.90</td>
<td>260.20</td>
<td>177.78</td>
<td>192.16</td>
<td>205.60</td>
<td>175.36</td>
</tr>
<tr>
<td>2</td>
<td>273.00</td>
<td>259.46</td>
<td>177.80</td>
<td>219.66</td>
<td>181.64</td>
<td>197.76</td>
</tr>
<tr>
<td>3</td>
<td>287.10</td>
<td>271.66</td>
<td>222.10</td>
<td>251.18</td>
<td>222.48</td>
<td>227.76</td>
</tr>
<tr>
<td>4</td>
<td>292.36</td>
<td>279.50</td>
<td>167.96</td>
<td>219.14</td>
<td>167.58</td>
<td>155.20</td>
</tr>
<tr>
<td>5</td>
<td>119.08</td>
<td>120.92</td>
<td>120.32</td>
<td>119.00</td>
<td>121.86</td>
<td>134.72</td>
</tr>
<tr>
<td><strong>Ave</strong></td>
<td>229.58</td>
<td>221.92</td>
<td>172.26</td>
<td>193.68</td>
<td>180.50</td>
<td>177.85</td>
</tr>
</tbody>
</table>

*Note.* GLIN = general linear classifier; GQC = general quadratic classifier; GCM(D) = generalized context model (deterministic); QNET = Quartic Network model; GDQC = general double quadratic classifier; QUAR = quadratic polynomial decision boundary; Ave = mean Akaike's information criterion value. Numbers in parentheses refer to the number of free parameters.

vertical linear boundary throughout the experiment. To summarize, simple visual inspection of the participants' response patterns suggests that for essentially all of the participants, models based on simple linear or quadratic decision boundaries will be inadequate.

**Model comparisons.** We fitted the GCM(D) and the GQC to the last 300 trials of the experiment for each participant, using the same fitting technique as in the preliminary model comparisons. The GCM(D) was fit using a euclidean distance metric and an exponential decay similarity function. We also fitted a three-parameter linear version of the GQC to the same data. In this model, the three quadratic terms in the decision bound \( h(x_1, x_2) \) are held fixed at zero, leaving two freely varying decision bound coefficients. Also, the effects of perceptual and criterial noise become nonidentifiable under this restriction, leaving only one variability parameter \( (\sigma_p + \epsilon)^2 \).

The best fitting AIC values for each model are listed in the first three columns of Table 4. The GCM(D) significantly outperformed the linear and full versions of the GQC for all participants (except Participant 5 in Condition 2). The average improvement of the GCM(D) over the full GQC was 49.66 AIC points. The magnitude of this difference provides convincing evidence against the use of a single linear or quadratic decision boundary in this task and also provides evidence that supports the exemplar storage and comparison processes underlying the GCM(D).

**Use of complex decision boundaries.** Although the preceding analyses provide evidence against the use of a single linear or quadratic decision boundary, there are several alternative hypotheses within the framework of decision-bound theory that deserve attention. One such hypothesis is that participants might have adopted linear or quadratic boundaries early in the task, but as they gained experience and realized that such boundaries yielded suboptimal performance, they then switched strategies. The plausibility of this hypothesis is diminished when one recalls that the most accurate linear boundary predicts performance that is only 5% below that predicted by the optimal likelihood boundary. Nevertheless, we tested this hypothesis by fitting the linear and full version of the GQC and the GCM(D) to Trials 101 through 400 of the first experimental session for each participant. All three models provided poor and roughly equivalent accounts of the data. Visual inspection of the data revealed highly noisy response patterns with no obvious interpretation. Possibly, participants adopted linear or quadratic boundaries and continuously adjusted their location in space early in the experiment in an attempt to learn the structure. Alternatively, highly noisy response patterns would be predicted by versions of the GCM(D) that introduced assumptions about probabilistic storage of exemplars and shifting patterns of selective attention to dimensions during early stages of learning. The early learning data are simply not highly diagnostic with respect to distinguishing between the alternative models.\(^5\)

\(^5\) We also fitted the GCM(D) to Trials 101–400 of the Maddox and Ashby (1993) data sets listed in Table 1. Maddox and Ashby had reported a substantial advantage for the GQC over the GCM(D), averaging over 20 AIC points across their 24 data sets. Our analyses, however, indicate that despite the fact that Maddox and Ashby used combinations of grid search and hill-climbing techniques, most of their results for the GCM(D) were local minima. After refitting their data, we find that the average advantage for the GQC over the GCM(D) in these data sets is only 6.47 AIC points. Much of this advantage is due to a single participant (Participant C2 from Data Set 3, Table 4, of Maddox & Ashby, 1993). If the fits for this participant are removed, the average advantage of the GQC over the GCM(D) is only 4.07 AIC points. As was the case for our first-session data, the overall fits of both models are relatively poor, reflecting the very noisy behavior observed during the early stages of learning.
A second avenue that needs to be explored is that rather than adopting linear or quadratic boundaries, participants adopted highly complex decision boundaries. To pursue this possibility, we developed and tested several alternative models positing highly complex decision bounds. We need to emphasize at the outset that we consider our investigation of these alternative models to be highly exploratory in nature and believe that the results should be interpreted with a good deal of caution. Decision bound theory provides a very general framework in which an infinite variety of alternative models can be expressed. As argued in our introduction, a decision bound can be essentially anything, and decision boundaries with alternative functional forms define entirely different classes of models. It is undoubtedly the case that good quantitative fits can be achieved for any regular subset of classification data by searching among the infinite variety of decision bounds and finding one that works, but such a procedure is of limited psychological interest.\(^6\)

Nevertheless, although the results should be interpreted with caution, we believe that the present explorations of alternative decision bounds are worthwhile for several reasons. First, they can provide converging evidence that participants in the present experiments were not constrained by the use of linear and quadratic boundaries. Second, they can provide a yardstick by which the performance of the exemplar-based GCM(D) can be evaluated. Third, they can point to promising directions for the development of alternative process models that give rise to complex decision boundaries. Because our investigations of these complex, high-parameter decision boundary models are exploratory, and are also secondary to our central goal of comparing the GCM(D) and the GQC, we describe the models in depth in the Appendix. The first model assumes that the participant uses two quadratic boundaries instead of a single one, and we call it the general double quadratic classifier (GDQC); the second model assumes that participants adopt a complex polynomial boundary of the fourth degree, and we call it the quartic polynomial decision boundary (QUARB); and the third model is a multilayered network that learns to implement a quartic polynomial decision boundary, and we call it the QuarNet model.

**Complex decision boundary model comparisons and discussion.** We fitted the GDQC, QUARB, and QuarNet models to the asymptotic response data for each participant. To provide some insight into the types of decision boundaries predicted by these models, Figure 5 shows the best fitting asymptotic decision boundaries predicted by each model for Participant 4 in Condition 1, along with the responses made during the last 300 trials of the experiment. Although these boundaries have noticeable differences, all display the same general shape and also do an excellent job of separating the Category A and B responses. Interestingly, all three models produce boundaries that accurately predict stimuli in the lower right quadrant to be classified into Category B.

The AIC values for the model fits are listed in Table 4 with those of the other models. Of the three complex decision boundary models tested, the general quartic model (QUARB) provided the best overall account of the data, with an average improvement in fit over the other two models of approximately 8 AIC points. More important, all three models provided better fits than the general quadratic classifier, each providing an average improvement in fit of at least 28 AIC points. These results provide additional evidence that participants' classification judgments were not constrained by linear or quadratic bounds. It is also the case that the GCM(D) provides better AIC fits, on average, than any of the complex decision boundary models. A better absolute fit (in terms of log likelihood) was provided by the QUARB model, but at the expense of using 15 free parameters. Although the AIC fit comparisons should be interpreted with caution, a reasonable conclusion is that the exemplar model provides a parsimonious process-level account of the underlying psychological basis for the best fitting complex decision boundaries. In Experiment 2 we investigate the generality of these results by comparing the models in an even more complex GRRRT category structure. 

**Experiment 2**

In Experiment 1, a single, relatively simple continuous curve (roughly cubic in form) could be placed between the category distributions such that a participant using this curve would correctly classify a large majority of the exemplars. An interesting theoretical question is whether natural categories display this property of separability by a simple continuous curve in perceptual space. If most natural categories do display this property, then it may be inherently difficult for people to learn category structures that are not easily separable in this manner. The purpose of Experiment 2 was to investigate this issue and also compare the models in a GRRRT category structure in which the optimal classification boundary is both nonquadratic and not easily describable in terms of a relatively simple continuous curve.

The category structure used in Experiment 2 is illustrated in Figure 6. The category distribution parameters are listed in Table 5. Category A is now a mixture of four normal densities, and Category B is a mixture of two densities. A convenient way of thinking about this category structure is that Category A exemplars tend to have extreme values on both stimulus dimensions, whereas Category B exemplars do not. Also illustrated in Figure 6 is the optimal likelihood classification boundary. Clearly, this boundary is both glo-

\(^6\) One can argue that because it is a six-parameter model, the GCM(D) also has a good deal of flexibility in fitting data and that numerous different versions of the GCM(D) are available, depending on which parameter constraints are used. Our judgment is that the flexibility that arises in the GCM(D), however, is of an entirely different kind than that which arises for decision bound theory. It is one thing to produce alternative versions of a model by allowing different parameter settings, but it is an entirely different thing to allow completely different decision boundaries from one application to another. The latter practice involves the application of entirely different classes of models. These differences are qualitative and structural in nature and are not the same type that arise from quantitative variation in parameter settings.
Figure 5. The best fitting decision boundaries for each of the complex decision boundary models for Participant 4 in Condition 1 of Experiment 1: (A) the general double quadratic classifier model, (B) the quadratic polynomial decision boundary model, and (C) the Quartic Network model. Also plotted are the last 300 responses made by Participant 4.

bally highly nonquadratic and not easily describable in terms of a simple continuous curve. A participant using the optimal likelihood classification boundary in this structure would correctly classify approximately 86% of the stimuli. By comparison, a participant using the most accurate linear boundary would correctly classify only 62% of the stimuli, and a participant using the most accurate quadratic boundary would correctly classify only 73% of the stimuli. Thus, it seems clear that if participants are to approach optimum in this task by using decision boundaries, they will be forced to adopt boundaries of a significantly more complex form than predicted by the GQC.

Method

Participants. All participants were graduate students at Indiana University who were unfamiliar with the theoretical issues under investigation in this research. (Participant 4 was aware of the GRTT paradigm, but was unaware of the category structures used
and the particular theoretical issues under investigation.) A total of 11 students participated in the experiment, for which they were paid $5 per experimental session plus a bonus of $10 for each percentage point above 70% achieved during the last half of each day's session. Participants were paid in full on completion of the experiment.

Stimuli and procedure. The stimuli, method of presentation, and experimental procedure were identical to that used in Experiment 1.

Results and Theoretical Analyses

Qualitative results. Table 6 shows the percentage correct obtained by each participant during each experimental session. Only 4 participants equaled or exceeded the performance predicted by the most accurate quadratic boundary (73%), and 3—Participants 2, 3, and 6—failed to equal the performance predicted by even the most accurate linear boundary (62%). Thus, although the category structure of Experiment 2 has a fairly simple logical description, it appears to have been relatively difficult for many participants to learn. In fact, the difficulty experienced by Participants 1 through 5 prompted us to test Participants 6 through 11 to obtain further evidence concerning the generality of this finding. Participants 6 through 11 were given the supplemental instruction that if they felt their performance was asymptoting at or below 73% by the end of the third session, they should make a concerted effort to improve beyond that point. This instruction appears to have had little impact on their overall performance.

The final session data of each of the 4 participants who exceeded the performance predicted by the most accurate quadratic boundary reflected the actual category distributions quite well, as is illustrated for Participant 4 in Figure 7A. The other 3 top-performers displayed very similar response patterns. The response patterns of the nonlearners, however, appeared to be quite variable and are difficult to describe succinctly. The data for Participant 10 are displayed in Figure 7B. This participant appears to have focused primarily on the angle dimension. More important, although a quadratic boundary does a good job of enclosing

<table>
<thead>
<tr>
<th>Category</th>
<th>g(x, y)</th>
<th>μ_x</th>
<th>μ_y</th>
<th>σ_x</th>
<th>σ_y</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>160</td>
<td>80</td>
<td>20</td>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>80</td>
<td>160</td>
<td>20</td>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>160</td>
<td>160</td>
<td>20</td>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>120</td>
<td>120</td>
<td>10</td>
<td>50</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>120</td>
<td>120</td>
<td>50</td>
<td>10</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Note. g(x, y) = density function; μ = mean; σ = standard deviation; ρ = correlation. Subscript r refers to radius of circle; subscript a refers to angle above horizontal. Angle dimension values are given in “units,” where π/4 radians equals 120 units.
Table 6

Percentage Correct Over Each Experimental Session for Each Participant in Experiment 2

<table>
<thead>
<tr>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>68</td>
<td>64</td>
<td>66</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>55</td>
<td>58</td>
<td>59</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>55</td>
<td>62</td>
<td>62</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>61</td>
<td>69</td>
<td>70</td>
<td>69</td>
<td>73</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>60</td>
<td>64</td>
<td>67</td>
<td>65</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>51</td>
<td>60</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>66</td>
<td>69</td>
<td>70</td>
<td>73</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>68</td>
<td>73</td>
<td>72</td>
<td>79</td>
</tr>
<tr>
<td>9</td>
<td>63</td>
<td>72</td>
<td>73</td>
<td>69</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>57</td>
<td>59</td>
<td>66</td>
<td>68</td>
<td>69</td>
</tr>
<tr>
<td>11</td>
<td>55</td>
<td>53</td>
<td>65</td>
<td>63</td>
<td>67</td>
</tr>
</tbody>
</table>

the Category B responses for most stimuli with an angle value near 120 units, there are a few noticeable exceptions to this boundary at extreme values of angle. The classification data for this participant (and also for Participant 7) turn out to be fairly well described by the GCM(D), with essentially all attention weight placed on a single dimension. Figure 7C shows the asymptotic response data for Participant 5. This participant appears to have focused primarily on the size dimension and may indeed have adopted a boundary that was roughly quadratic. The response patterns for Participants 6 and 11 display a somewhat similar pattern. Finally, Figure 7D shows the response data for Participant 3, one of the very poor performing participants. This participant’s asymptotic response pattern was extremely noisy, and we have been unable to provide an interpretation for the data. The response pattern for Participant 2 was also somewhat noisy and difficult to interpret.

Model comparisons. We fitted the GQ, GCM(D), and complex decision boundary models to the last 300 trials of the experiment for each participant. The AIC fits are listed in Table 7. The first main result is that the GCM(D) outperformed the GQ for all but 2 participants (Participants 5 and 6). The GQ performs poorly for a variety of reasons. First, because the response patterns of the top-performing participants approximated the underlying category distributions, and because the boundary partitioning of these categories is highly nonquadratic, a single quadratic decision boundary simply cannot describe these participants’ data. Second, the poor performance of the GQ in predicting the data can also be explained by referring to the responses made by Participant 10 shown in Figure 7B. Recall that this response pattern is not well approximated by a single quadratic boundary because of exceptions to such a boundary at extreme values of angle. The best fitting GQ boundary for Participant 10 is plotted along with the data in Figure 8. This boundary, although capturing many of the participant’s responses, strongly mispredicts these exceptions, resulting in a poor fit. In summary, direct comparisons between the GCM(D) and the GQ for the present category structure strongly favor the GCM(D).

When compared with the complex decision boundary models, the overall fit of the GCM(D) is much less impressive. In favor of the GCM(D) is the fact that in terms of AIC, it is the best fitting model for 3 of the 4 top-performing participants (1, 4, and 8) and comes close to being the best fitting model for Participants 7 and 10, who are next in line in order of performance (see Tables 6 and 7). When attention is restricted to the 6 top-performing participants, the average AIC for the GCM(D) surpasses that of any of the complex decision boundary models. However, for those participants who performed poorly, the fit of the GCM(D) tends to be quite poor. Averaged across all participants, the model that assumes double quadratic boundaries (the 12-parameter GDQC) provides the best AIC fit.

In a nutshell, then, when participants did fairly well at learning the category structures, the GCM(D) did a relatively good job of predicting their performance. Thus, for the top-performing participants, it again appears that the GCM(D) provides a reasonable process-level interpretation of the pattern of results. However, when participants did not learn the category structures, the performance of the model was poor.

This general pattern of results is not surprising in light of what is known about the GCM. With an infinite number of training exemplars and with an infinite value of the sensitivity parameter (plus at least some attention given to each dimension), the equivocality contour of the GCM matches the optimal decision boundary. With fewer numbers of training exemplars and finite sensitivity, the equivocality contour of the GCM approximates the optimal boundary (e.g., Maddox & Ashby, 1993). The GCM basically predicts, therefore, that individuals will learn the category distributions regardless of their complexity. The model runs into trouble in situations in which highly suboptimal response patterns are observed and in which the participants’ responses do not approximate the category distributions. In these situations, high-parameter models that allow for a great deal of flexibility in the shapes of the decision boundaries tend to have an advantage. This flexibility is certainly inherent in the complex decision boundary models tested here.

General Discussion

The main focus of this research was to compare and contrast an exemplar model, the GCM(D), and a recently proposed decision boundary model, the GQ, on their ability to predict asymptotic, individual participant response data in a series of classification experiments involving large, ill-defined category structures. These experiments involved designs in which participants experienced thousands of unique training exemplars and in which the optimal classification boundary for separating the members of the alternative categories was highly nonlinear. Because most previous successes of the exemplar model have been obtained in designs using relatively small-sized category structures, it is of fundamental interest whether the exemplar model remains competitive under the present conditions. The experiments are also of fundamental interest from the perspective of decision bound theory, because they provide highly
diagnostic information regarding whether individuals are predisposed or constrained to use simple linear or quadratic decision boundaries.

Our direct quantitative comparisons between the exemplar model, the GCM(D), and the GQC provided some valuable information. First, in preliminary comparisons, we established that the superiority of the GQC reported by Maddox and Ashby (1993) in structures with quadratic optimal boundaries was largely an artifact of the approximation fit procedure used in that study. When the true GCM(D) was fit to the data, neither model displayed a distinct quantitative advantage over the other. Furthermore, we advanced theoretical arguments that even if the GCM(D) did, in fact, come close to being the “true” model, it would be difficult to demonstrate a distinct superiority for the GCM(D) over the GQC for the category structures considered by Maddox and Ashby. The reason is that when the categories are multivariate normal, the “boundary” predicted by the GCM(D) is always roughly quadratic in form, with parameters highly constrained by the set of training

Figure 7. The last 300 responses made by (A) Participant 4, (B) Participant 10, (C) Participant 5, and (D) Participant 3, in Experiment 2. Crosses refer to Category A responses, and circles refer to Category B responses. Panel A also shows the optimal classification boundary.
exemplars. By contrast, the GQC is free to predict any quadratic boundary and, thus, has considerably more flexibility than the GCM(D) when the categories are multivariate normal.

Thus, in an effort to develop more diagnostic tests, we designed category structures composed of mixtures of multivariate normal distributions. For all of these structures, the category “boundary” predicted by the GCM(D) was highly nonquadratic, thereby allowing for sharp contrasts to be established between the GCM(D) and the GQC. The results of these experiments provided strong evidence against the use of a single quadratic decision boundary. For virtually all participants in both experiments, the GCM(D) yielded far better fits to the asymptotic classification data than did the GQC. We obtained converging evidence against the use of simple linear or quadratic decision bounds by demonstrating that various complex decision bound models also yielded far better fits to the classification data than did the GQC. Finally, the results of these model comparisons were corroborated by simple visual inspection of the asymptotic response patterns of the individual participants. Because a simple linear or quadratic boundary would have yielded performance nearly as good as the optimal boundary in Experiment 1, the finding that observers failed to use such boundaries in that experiment seems quite damaging to the theory. These results provide a significant challenge, therefore, to the quadratic classifier as a general model of human classification.

In the present experiments, the category structures were composed of mixtures of bivariate normal distributions. One approach to extending the quadratic decision bound model might involve the assumption that during the course of learning, participants detected that the categories were mixtures, and then engaged in a process of subcategorizing. For example, in Experiment 1, each superordinate category,

A and B, was composed of two subordinates, A1 and A2, and B1 and B2, respectively. A participant might have formed decision boundaries to subdivide the perceptual space into these four subregions. Then, the subcategorizations would be combined to yield the overall categorization response. For example, anytime a participant classified an object into A1 or A2, a Category A response would be made. The optimal boundaries for dividing the space into the four regions corresponding to each bivariate normal are quadratic in form. Indeed, if these quadratic subboundaries are combined, they then yield the optimal boundary for separating the members of the two superordinate categories, A and B. This subcategorization hypothesis is closely related to an intriguing decision bound hypothesis proposed by Maddox and Ashby (1993), namely, that in any given classification paradigm, participants will adopt a boundary with the same functional form as the optimal boundary.\footnote{In its original statement (Ashby, 1992; Ashby \& Maddox, 1992), the hypothesis seemed to be that because many natural categories are normally distributed and because the optimal classification boundary separating multivariate normals is quadratic in form, participants have learned through past experience to use quadratic boundaries and apply them across diverse tasks. Maddox and Ashby’s (1993) more recent statement of the hypothesis, however, might lead them to disavow application of the GQC for the category structures tested in this study, in which the optimal decision boundaries were nonquadratic in form. We cannot be sure, because Maddox and Ashby (1993, pp. 60–67) tested and successfully applied the GQC in other contexts in which the form of the optimal decision boundary was nonquadratic, so its domain of intended applicability remains unclear at this stage. Another difficulty is that except for category structures defined over certain special mathematical distributions (e.g., the multivariate normal),

### Table 7

**Model Fits (Akaike’s Information Criterion) to Last 300 Trials of Final Experimental Session in Experiment 2**

<table>
<thead>
<tr>
<th>Participant</th>
<th>GQC (7)</th>
<th>GCM(D) (6)</th>
<th>ONET (10)</th>
<th>GDQC (12)</th>
<th>QUIARB (15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>327.34</td>
<td>210.14</td>
<td>294.82</td>
<td>217.60</td>
<td>231.40</td>
</tr>
<tr>
<td>2</td>
<td>304.34</td>
<td>303.82</td>
<td>274.90</td>
<td>269.44</td>
<td>271.86</td>
</tr>
<tr>
<td>3</td>
<td>396.20</td>
<td>373.54</td>
<td>372.28</td>
<td>350.10</td>
<td>347.38</td>
</tr>
<tr>
<td>4</td>
<td>328.38</td>
<td>218.72</td>
<td>236.50</td>
<td>225.20</td>
<td>240.68</td>
</tr>
<tr>
<td>5</td>
<td>192.20</td>
<td>231.46</td>
<td>181.86</td>
<td>182.86</td>
<td>195.04</td>
</tr>
<tr>
<td>6</td>
<td>293.42</td>
<td>308.60</td>
<td>308.76</td>
<td>279.62</td>
<td>265.56</td>
</tr>
<tr>
<td>7</td>
<td>254.70</td>
<td>199.26</td>
<td>222.06</td>
<td>195.08</td>
<td>202.82</td>
</tr>
<tr>
<td>8</td>
<td>321.46</td>
<td>173.82</td>
<td>209.78</td>
<td>181.20</td>
<td>174.64</td>
</tr>
<tr>
<td>9</td>
<td>313.74</td>
<td>204.44</td>
<td>215.04</td>
<td>197.86</td>
<td>219.50</td>
</tr>
<tr>
<td>10</td>
<td>299.74</td>
<td>228.50</td>
<td>242.52</td>
<td>227.24</td>
<td>241.86</td>
</tr>
<tr>
<td>11</td>
<td>236.92</td>
<td>234.78</td>
<td>272.90</td>
<td>213.54</td>
<td>202.94</td>
</tr>
<tr>
<td>Ave</td>
<td>297.13</td>
<td>244.28</td>
<td>257.40</td>
<td>230.88</td>
<td>245.36</td>
</tr>
</tbody>
</table>

**Note.** GQC = general quadratic classifier; GCM(D) = generalized context model (deterministic); QNET = Quartic Network model; GDQC = general double quadratic classifier; QUIARB = quadratic polynomial decision boundary; Ave = mean Akaike’s information criterion. Numbers in parentheses refer to the number of free parameters.

---

**Figure 8.** The best fitting general quadratic classifier decision boundary for Participant 10 in Experiment 2, along with the final 300 responses.
Although we cannot rule out the subcategorization hypothesis on the basis of the present data, several factors argue against the utility of the idea. First, like the other complex decision boundary models that we considered in this article, a model that posits combinations of quadratic curves would require that a large number of free parameters be estimated. Various post hoc assumptions would need to be introduced to reduce the number of free parameters so as to allow for parsimonious fits to the data. Second, the subcategorization hypothesis seems to beg the question of how participants form decision boundaries in the first place. The assumption that participants have knowledge of the underlying distributional structure of the categories seems to presuppose some type of exemplar storage and comparison process, precisely as assumed in the GCM. Indeed, taken to the limit, the subcategorization hypothesis seems very much like an exemplar model. If the categories are composed, say, of mixtures of numerous different exemplar distributions, would the decision boundary theorist then posit a distinct quadratic boundary to subclassify each and every distribution? The critical point is that for the subcategorization hypothesis to have theoretical utility, the decision boundary theorist needs to specify the experimental conditions that give rise to multiple boundaries and predict them a priori.

Finally, we have conducted some preliminary empirical tests of the subcategorization hypothesis and thus far have found no evidence to support it. Using 5 new participants, we replicated Condition 1 of Experiment 1. On the final day of testing, however, we provided participants with explicit information that Categories A and B were both composed of two subcategories. The participants were given the following instructions. First, they were to make their main categorization response, A or B, just as they had been doing all along. Next, they were to free-classify the object into one of the two appropriate subcategories. We reasoned that if participants had recognized that the categories were composed of mixtures of normals and had been subclassifying all along, then their free classifications should trace a boundary that was fairly close to the optimal one for separating the space into the four subregions. The results proved otherwise, however. Instead, all of the participants apparently used a decision rule in which they selected a particular value along one of the dimensions and used this value as a dividing line for their subclassifications. For example, after classifying an object into Category A, a participant might place it in A1 if its size exceeded x units and into A2 otherwise (see Ahn & Medin, 1992, and Medin, Wattenmaker, & Hampson, 1987, for similar results concerning the nature of free classification strategies). This type of decision boundary is much different in form from the optimal boundaries for separating the four bivariate normals, which show extreme quadratic curvature. It is also worth noting that in interviews conducted following the experiment, all of the participants indicated that they had no idea that the main categories were composed of subcategories, and all expressed surprise regarding the subcategorization instructions provided on the final day of testing. Despite these problems associated with the subcategorization hypothesis, we find the idea intriguing, and it is certainly worthy of more rigorous and continued investigation.

Although most of our experimental results were consistent with the exemplar-storage processes posited by the GCM (and suggested that participants can learn very complex decision boundaries), some of the results appear to be quite damaging to exemplar theory. In particular, the Experiment 2 structure proved to be quite difficult for many of our participants to learn. The GCM(D) did an excellent job of predicting the data for the participants who were able to learn the task. It also yielded good fits to the data of some nonlearners who appeared to focus almost all of their attention on a single stimulus dimension. However, the model did a relatively poor job of predicting the data for the very poor performing participants, and various alternative complex decision boundary models yielded much better fits to the data. The GCM(D) failed to yield adequate fits to these data because it basically predicts that given enough experience with training exemplars, participants' response patterns should eventually approximate the underlying category distributions. The complex decision boundary models, on the other hand, make only minimal predictions concerning the learnability of a particular task and have a great deal more flexibility for predicting highly suboptimal patterns of performance. Thus, they did a better job of fitting the Experiment 2 data for the participants who did not learn the category distributions.

On the one hand, the ability of the GCM(D) to accurately predict the asymptotic response patterns for most of the individual participants in these experiments strikes us as fairly impressive. Unlike previous tests involving small-size, deterministic structures, the present tests involved the presentation of thousands of training exemplars that were probabilistically assigned to ill-defined categories. Thus, the successes of the model in the present design represent a significant demonstration of the potential generalizability of exemplar-based approaches.

On the other hand, the failure of roughly half the participants to learn the category structure of Experiment 2 suggests possible limitations in the ability of humans that are not well predicted by the exemplar storage and comparison processes underlying the GCM(D). Of course, one could argue that had classification training continued indefinitely, all participants would eventually have learned the structure. However, the present task involved highly motivated participants, who, after a full week of training, had experienced 4,000 category exemplars. Thus, at the very least, the results suggest a practical limit on humans' classification abilities.\(^8\)

\(^8\) The range of stimulus values and the populations from which participants were recruited were essentially identical across Experiments 1 and 2. Therefore, the values of the model parameters should have been nearly the same across the two experiments. Given the range of parameter estimates obtained in Experiment 1, the GCM(D) predicts that all participants should have learned the category structures in Experiment 2.
Humans, it appears, are not predisposed to learn arbitrarily complex category structures. A possible explanation for this inability could be that the learning of natural categories does not require the use of arbitrarily complex boundaries (Ashby, 1992). Unfortunately, we presently know very little about how natural categories are distributed along psychological dimensions, so it is difficult to evaluate the merits of this hypothesis. Nonetheless, the inability of many participants to learn the structure of Experiment 2 presents a considerable challenge to the exemplar approach to modeling classification and also an important direction for future research and model development.

References


EXEMPLAR AND DECISION BOUND MODELS


Appendix

Complex Decision Boundary Models

In this Appendix, we describe in detail the three complex decision boundary models that we formulated and tested in Experiments 1 and 2.

General Double Quadratic Classifier (GDQC)

The first model assumes that the participant uses two quadratic decision boundaries instead of a single quadratic boundary. This model is based on the idea that if categorizers have extensive experience working with quadratic decision boundaries, then they may be predisposed to use boundaries of that form, even in tasks in which one such boundary is inadequate. Thus, in more complex tasks such as that in Experiment 1, categorizers may adopt a pair of quadratic boundaries.

One difficulty in developing a model of this type involves deciding how to combine the quadratic decision regions from the two boundaries to form the response region for a particular category. We investigated versions of the model that identify the Category A response region as intersection of the two quadratic regions and also as the union of the two regions. The version of the model that used the intersection of the regions provided consistently better fits, so we focus on that version. Formally, the model uses a pair of independent quadratic boundaries, \( h_1(x_1, x_2) \) and \( h_2(x_1, x_2) \), and gives the probability of a Category A response on trials in which stimulus \( i \) was presented as the probability that both \( h_1 \) and \( h_2 \) exceed some criterion \( \delta \). That is,

\[
Pr[A|i] = Pr[h_1(x_{pi}) + \varepsilon > \delta_1] \cdot Pr[h_2(x_{pi}) + \varepsilon > \delta_2].
\]  

(A1)

where the same procedures as described previously are used in computing the component probabilities. Criterial variability is assumed to be the same for both decision bounds. Thus, the GDQC model uses 12 free parameters: five freely varying decision bound parameters for each decision boundary, a perceptual variability parameter (\( \sigma^2 \)), and a criterial variability parameter (\( \sigma_c^2 \)). The two bias parameters (\( \delta_1 \) and \( \delta_2 \)) were held fixed at zero.

Quartic Polynomial Decision Boundary Model (QUARB)

Another possibility is that participants simply abandon the use of quadratic boundaries when they fail to yield adequate performance, and instead adopt a boundary represented by a higher order polynomial. Note that any analytic function can be well approximated by using a polynomial of sufficient degree, so this model has considerable flexibility. The psychological basis for the use of a higher degree polynomial seems questionable to us, but we decided to explore such a model nevertheless. The model we developed assumed that the participant enters into the task with a quartic-level polynomial decision boundary. (A cubic-level model was also investigated but provided worse AIC fits than the quartic-level model.) With two stimulus dimensions, this model assumes the participant uses a decision boundary of the form

\[
u(x_1, x_2) = a_1 \cdot x_1^4 + a_2 \cdot x_1^3 \cdot x_2 + a_3 \cdot x_1^2 \cdot x_2^2 + \ldots + a_{13} \cdot x_1 + a_{14} \cdot x_2 + a_{15}
\]

(A2)

and assigns the probability of a Category A response on trials in which stimulus \( i \) is presented as

\[
Pr[A|i] = Pr[u(x_{pi}) + \varepsilon > \delta].
\]

(A3)

where \( \varepsilon \) is a random variable that reflects criterial variability. As in the GQC, \( \varepsilon \) is assumed to be normally distributed with mean zero and variance \( \sigma_c^2 \). We have assumed that there is no perceptual variability, thus \( u(x_{pi}) \) is not a stochastic quantity. This assumption is needed because for high-degree polynomials, \( u(x_{pi}) + \varepsilon \) may not be well approximated by a normal distribution when perceptual variability is high. The assumption of accepting zero perceptual variability seems reasonable for the present response-terminated visual displays. The QUARB model has a total of 15 free parameters: 14 freely varying decision bound parameters and a criterial variability parameter (\( \sigma_c^2 \)).

Quartic Decision Bound Network Model (QuarNet)

The freely varying quartic polynomial of the QUARB model gives it extreme flexibility in fitting participant data. It pays for this flexibility, however, with its large number of free parameters. An important research direction for models developed from decision bound theory is to address the process by which a particular boundary develops as a function of experience. As a preliminary step in this direction, we developed and tested a decision boundary learning model known as QuarNet (for Quartic Network; see Bussemeyer & Myung, 1992, for alternative possibilities).

The intuition underlying the QuarNet model is that the participants enter classification learning tasks with very flexible decision boundaries at their disposal and adjust the position of these boundaries in perceptual space on a trial-by-trial basis in an attempt to minimize categorization errors. Formally, the model is instantiated in a three-layer feed-forward connectionist network, illustrated in Figure A1. Although the model can be generalized in a straightforward manner to situations involving any number of stimulus dimensions and categories, the description of the model that fol-
lows assumes a structure with two stimulus dimensions and two categories.

At the input layer, the model codes the physical stimulus values, \( x_1 \) and \( x_2 \), and allows for a freely varying affine transformation of these values into psychological values, \( x_{p1} \) and \( x_{p2} \), given by

\[
x_{pi} = A_i \cdot x_i + B_i, \quad \text{for } i = 1, 2.
\]  
\[(A4)\]

These psychological values are then passed to the hidden layer, where each hidden node computes a unique term in the general quartic equation in two dimensions. That is, the activation of the 1st hidden node is given by \( a_{h1} = x_{p1}^4 \), the activation of the 2nd hidden node is \( a_{h2} = x_{p1}^3 \cdot x_{p2} \), and so on. The activation of the 15th hidden node, which computes the constant term in the equation, is fixed at \( a_{h15} = 1 \). These hidden node activations are weighted and passed to the category output nodes, where they are summed to produce strength values for the alternative categories. Thus, the activation of output node \( k \) is

\[
O_k = \sum w_{ik} \cdot a_{hi}.
\]  
\[(A5)\]

The weights \( w_{ik} \) can be thought of as coefficients in the quartic decision boundaries used for each category. These output activations are converted into response probabilities, using the logistic transformation

\[
Pr[A] = \frac{\exp(c \cdot O_1)}{\exp(c \cdot O_1) + \exp(c \cdot O_2)}.
\]  
\[(A6)\]

where \( c \) is a freely varying scaling constant. This response function can be viewed as a deterministic response mechanism, analogous to that used in the QUARB model, which assumes logistically distributed criterial noise.

All weights in the network are initialized at zero. Learning of the weights occurs by using the delta rule ( Widrow & Hoff, 1960). When Category A feedback is presented, the Category A output node receives a teaching signal of \( t_A = 1 \) and the Category B output node receives a teaching signal of \( t_B = -1 \). When Category B feedback is given, the teaching signals are reversed. All weights in the network are then updated by the rule

\[
\Delta w_{ik} = \beta M \cdot (t_k - O_k) \cdot a_{hi}.
\]  
\[(A7)\]

where \( \beta_M \) is a learning rate parameter for all nodes that compute terms of order \( M \). That is, there is a learning rate \( \beta_M \) for the node that computes the constant term, a learning rate \( \beta_1 \) for the nodes that compute the \( x_1^1 \) and \( x_2^1 \) terms, and so on. The QuarNet model has a total of 10 free parameters: five learning rate parameters \( \beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \), 2 scale parameters \( A_1 \) and \( A_2 \) and 2 translation parameters \( B_1 \) and \( B_2 \) used in the affine transformation of the physical stimulus values, and a scaling parameter \( c \) used in converting category node outputs into response probabilities.