Psychonomic Bulletin & Review
1998, 5 (3), 345-369

A rule-plus-exception model for classifying objects in continuous-dimension spaces

ROBERT M. NOSOFSKY
Indiana University, Bloomington, Indiana

and

THOMAS J. PALMERI
Vanderbilt University, Nashville, Tennessee

The authors propose a rule-plus-exception (RULEX) model for how observers classify stimuli residing in continuous-dimension spaces. The model follows in the spirit of the discrete-dimension version of RULEX developed by Nosofsky, Palmeri, and McKinley (1994). According to the model, observers learn categories by forming simple logical rules along single dimensions and by remembering occasional exceptions to those rules. In the continuous-dimension version of RULEX, the rules are formalized in terms of linear decision boundaries that are orthogonal to the coordinate axes of the psychological space. In addition, a similarity-comparison process governs whether stored exceptions are used to classify an object. The model provides excellent quantitative fits both to averaged classification transfer data and to distributions of generalizations observed at the individual-participant level. The modeling analyses suggest that, when multiple rules are available for solving a problem, averaged classification data often represent a probabilistic mixture of idiosyncratic rule-plus-exception strategies.

The idea that people may represent categories in terms of simple logical rules dates back to the very beginnings of research on concept identification in cognitive psychology (Bourne, 1970; Bruner, Goodnow, & Austin, 1956; Hunt, Marin, & Stone, 1966; Levine, 1975; Restle, 1962; Trabasso & Bower, 1968). The rule hypothesis carries a good deal of intuitive appeal. An important purpose of categorization is to reduce the complexity of mental processing by organizing distinct objects into classes and then dealing with the classes as wholes rather than with each object uniquely. By forming a simple rule, an economical summary description is provided for an entire class of objects, thereby allowing for a vast reduction in the amount of information that one needs to store in memory. Furthermore, to decide category membership for any individual object, one need only decide whether or not the combination of attributes that composes the object satisfies the rule.

Despite its intuitive appeal and the early dominance of this approach, models based on the formation of simple logical rules had, until recently, largely dropped from the scene in categorization research. Historically, the main impetus for this trend can be traced to the highly influential work of such researchers as Posner and Keele (1968) and Rosch (1973; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976). Rosch, for example, argued convincingly that most categories in the natural world were not structured according to simple rules. For one thing, it is difficult to state simple rules or definitions that perfectly partition the members of most natural-world categories. Furthermore, experimental research indicates that categories have a graded, internal structure, in which some objects are "better" or more typical members of the category than others are (Barsalou, 1985; Rips, Schoben, & Smith, 1973; Rosch, 1973). Models based solely on the idea that simple rules or definitions are used to represent categories seem unable to account for the full range of findings involving these typicality and graded-structure effects (for an extensive review and analysis, see E. E. Smith & Medin, 1982, chap. 3).

In response to these challenges, experimental research began to examine the learning of ill-defined category structures in which no simple rules existed for deciding category membership. A wide variety of models have been developed to account for the learning of such ill-defined categories. These models include prototype models (Reed, 1972), feature-set models (Hayes-Roth & Hayes-Roth, 1977), exemplar models (Medin & Schaffer, 1978), connectionist models (Knapp & J. A. Anderson, 1984), Bayesian models (J. R. Anderson, 1991), and decision-boundary models (Ashby & Lee, 1991). According to exemplar models, for instance, people represent categories by storing previously experienced category exemplars in memory and classifying objects on the basis of their similarity to these exemplars. Such models, which have enjoyed enormous success at accounting for diverse cate-

This work was supported by Grant PHS RO1 MH48494-06 from the National Institute of Mental Health. The authors would like to thank William Estes, John Kruschke, W. Todd Maddox, and Andre Vandierendonck for their helpful criticisms of an earlier version of this article. Correspondence concerning this article should be addressed to R. M. Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405 (e-mail: nosofsky@indiana.edu).

gorization phenomena (Brooks, 1978; Estes, 1994; Heit, 1994; Hintzman, 1986; Kruschke, 1992; Lamberts, 1994; Medin & Schaffer, 1978; Nosofsky, 1986), provide a view of category representation that is dramatically different in spirit from those of simple rule-based models.

Recently, however, the pendulum of competing ideas has begun to shift back to consideration of simple rule-based models, at least as an important component of category learning and representation. One reason for this renewed consideration is quite subjective in nature: People have strong impressions that they do indeed form rules when learning to categorize. Furthermore, some researchers have questioned the plausibility of exemplar memory models and the vast storage and computational resources that they seem to require.

More important, researchers have begun to develop elaborated rule-based models that are showing some promising successes at accounting for various forms of classification and memory data (Ahn & Medin, 1992; Martin & Caramazza, 1980; Medin, Wattenmaker, & Michalski, 1987; Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995; Ward & Scott, 1987). The most broadly tested of these models in the domain of classification learning is the rule-plus-exception (RULEX) model of classification proposed by Nosofsky, Palmeri, and McKinley (1994; Palmeri & Nosofsky, 1995). Following in the spirit of the early concept-identification models, according to RULEX, observers learn to classify by a process of hypothesis testing in which simple logical rules are formed along the dimensions that compose the objects. As a straightforward extension of such models, RULEX allows for the formation of imperfect rules—that is, rules that do not perfectly partition the members of contrasting categories. Observers then supplement these imperfect rules with occasional stored exceptions. Thus, complex, ill-defined categorization problems can often be solved by a combination of rule formation and exception storage. An important theme in the model is that large individual differences are expected to be observed in the particular rules that are formed and in the exceptions that are stored, so that averaged classification data may not be representative of the behavior of any single subject.

Nosofsky, Palmeri, and McKinley (1994) demonstrated that RULEX provides excellent quantitative accounts of a wide variety of benchmark phenomena in the modern categorization literature. These phenomena include prototype and specific exemplar effects (Medin & Schaffer, 1978), selective attention effects (Medin & E. E. Smith, 1981; Nosofsky, 1984), sensitivity to correlated dimensions (Medin, Altom, Edelson, & Freko, 1982), the difficulty of learning linearly versus nonlinearly separable categories (Medin & Schwanenflugel, 1981), and the relative difficulty of learning categorization problems described by rules of differing complexity (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard, Hovland, & Jenkins, 1961). Furthermore, Palmeri and Nosofsky (1995) provided converging evidence for the RULEX ideas by collecting recognition memory data following the completion of category learning. In these studies, observers displayed superior memory for exceptions to category rules, and a mixed model that assumed a combination of RULEX processing and residual exemplar storage provided good quantitative accounts of the recognition data.

An important limitation of RULEX, however, is that the current version predicts performance only for stimuli varying along discrete binary-valued dimensions. Although numerous categorization experiments have been conducted in such a stimulus domain, to evaluate the generality with which RULEX processing may occur, it is vital to extend the model to predict performance in continuous-dimension stimulus domains. Indeed, a common critique of the concept-identification paradigm and the associated hypothesis-testing models that were developed is that many real-world categories are composed of continuous rather than discrete dimensions (Reed, 1996, pp. 225–226). Therefore, the main purpose of this research was to begin the development and testing of a continuous-dimension RULEX model.

We organize our article as follows. We start by briefly reviewing how RULEX has been used to model classification learning for stimuli varying along binary-valued dimensions. Using this previous modeling as a guide, we then develop a continuous-dimension version of RULEX. Nosofsky, Palmeri, and McKinley (1994) implemented the binary-valued version of RULEX as an explicit learning model. However, because the complexity of modeling performance is far greater in multivalued, continuous-dimension domains than in binary-dimension ones, in the present article we bypass questions about learning and simply propose an asymptotic form of the model. This continuous-dimension RULEX model is then evaluated by fitting it to data from several previously published studies in the literature, as well as to some new data sets reported herein. For many of these studies, we believe that the continuous-dimension RULEX model provides an account of the complete set of data that is as good as or better than the accounts provided by extant alternative models. We conclude the article by briefly comparing RULEX with some related models that are currently being developed by other investigators. To forecast this final discussion, we argue that although RULEX processing may play a fundamental role in a variety of categorization settings, we do not view RULEX as a self-sufficient model. Rather, our ultimate aim is the development of a hybrid model that includes both RULEX processing and exemplar-based processing as fundamental components.

Review of the Binary-Valued RULEX Model

Consider the category structure shown in Table 1. This structure was used by Medin and Schaffer (1978) in their seminal article which introduced the exemplar-based *context model* and has been used by numerous investigators since then. The stimuli vary along four binary-valued dimensions. There are five Category A training exemplars, four Category B training exemplars, and seven transfer

Table 1
Example of Category Structure Tested in Medin and Schaffer's (1978) Experiments 2 and 3

Category A	Category B	Transfer Stimuli
A1 1112	B1 1122	T1 1221
A2 1212	B2 2112	T2 1222
A3 1211	B3 2221	T3 1111
A4 1121	B4 2222	T4 2212
A5 2111		T5 2121
		T6 2211
		T7 2122

stimuli. Logical Value 1 on each dimension tends to indicate Category A, and Logical Value 2 tends to indicate Category B, but there are no singly necessary and jointly sufficient sets of features that define the categories. According to RULEX, by the time the learning process is completed, an individual observer might have stored the following information in memory. First, the observer might store the (imperfect) single-dimension rule that objects with Value 1 on Dimension 1 belong to category *A* and that objects with Value 2 on Dimension 1 belong to Category B (see Table 1). We summarize these rules by using the notation 1*** → A, 2*** → B, where the asterisks denote dimension "wild cards" that match any value. Exemplars A5 and B1 are exceptions to this rule, so the observer must store additional information to learn the categories. For example, the observer might store the exceptions 2*11 → A, 1*22 → B (see Table 1). Note that, with these rules, the categorization problem is solved, even though no complete exemplars are stored in memory. The learning process in RULEX is stochastic, and a key property of the model is that different observers form alternative rules and exceptions. For example, numerous observers might, instead, form rules along Dimension 3, **1* → A, **2* → B, and store information to classify the A4 and B2 exceptions—for example, 1*21 → A, 2*12 → B. Averaged classification data are assumed to represent probabilistic mixtures of these idiosyncratic rules and exceptions. An explicit learning process is formalized in the RULEX simulation that incorporates classic principles of hypothesis testing (e.g., Levine, 1975; Trabasso & Bower, 1968) and probabilistic storage of exception information. Thus, although a vast array of different rules and exceptions are involved in predicting the averaged classification data, these rules and exceptions emerge from a probabilistic learning process described by relatively few free parameters.

Nosofsky, Palmeri, and McKinley (1994) demonstrated that RULEX provides excellent quantitative fits to averaged classification data, fits that are essentially the same as those achieved by the exemplar-based context model. Beyond predicting averaged classification data, however, RULEX also fares well at predicting patterns of performance at the individual observer level. A highly diagnostic form of data is what Nosofsky, Palmeri, and McKinley (1994) referred to as a *distribution of generalizations* (see also Nosofsky, Clark, & Shin, 1989; Pavel, Gluck, & Henkle, 1988). Consider the transfer stimuli in Table 1. During test, each transfer stimulus is classified by an individual observer into either Category A or Category B. The specific pattern of classification responses given to the transfer stimuli defines a *generalization profile* for an individual observer. For example, an observer classifying T1–T3 into Category A and T4–T7 into Category B yields the generalization profile AAABBBB. The distribution of generalizations is then obtained by computing the frequency of individuals displaying each profile. The top panel of Figure 1 shows the distribution of generalizations observed in Nosofsky, Palmeri, and McKinley's (1994) replication of Medin and Schaffer's (1978) experiment. The bottom panel shows the distribution predicted by RULEX. (This distribution was predicted while holding fixed the parameters that best fit the averaged transfer data, although in the present article we will have reason to fit both types of data simultaneously.) RULEX does a reasonably good job of predicting the observed distribution. This achievement is important, because in addition to accounting for the averaged transfer data, RULEX simultaneously characterizes the patterns of performance observed at the individual observer level. By contrast, the exemplar-based context model failed dramatically to predict the distribution-of-generalization data (see Nosofsky, Palmeri, & McKinley, 1994). In evaluating the continuous-dimension version of RULEX in this article, we continue to rely on distribution-of-generalization data to provide more incisive tests of the model.

A Continuous-Dimension RULEX Model

To introduce the continuous-dimension RULEX model, we refer to the category structure illustrated in Figure 2, which was tested in a previous study by Nosofsky et al. (1989). The stimuli were circles varying in size and angle of orientation of a radial line, which are highly separable dimensions (Nosofsky, 1985; Shepard, 1964). There were four levels of size and four levels of angle combined orthogonally to yield 16 stimuli. The spacings between dimension values illustrated in Figure 2 were derived in a separate similarity-scaling study. Stimuli enclosed by circles and triangles represent Category A and Category B training exemplars, respectively; the unenclosed stimuli were novel transfer items. Nosofsky et al. (1989) designed this category structure to contrast the predictions of the exemplar-based context model with those of a particular rule-based model formalizing an "economy-of-description" view. Although the results favored the context model over the economy-of-description rule model, we will see that the continuous-dimension version of RULEX fares even better at accounting for the data.

A natural way of extending RULEX to the domain of continuous-dimension stimuli is to make use of the *decision-boundary* construct central to the general recognition theory (GRT) of Ashby, Townsend, and their associates (e.g., Ashby & Gott, 1988; Ashby & Townsend, 1986; Maddox & Ashby, 1993). GRT is a multidimen-

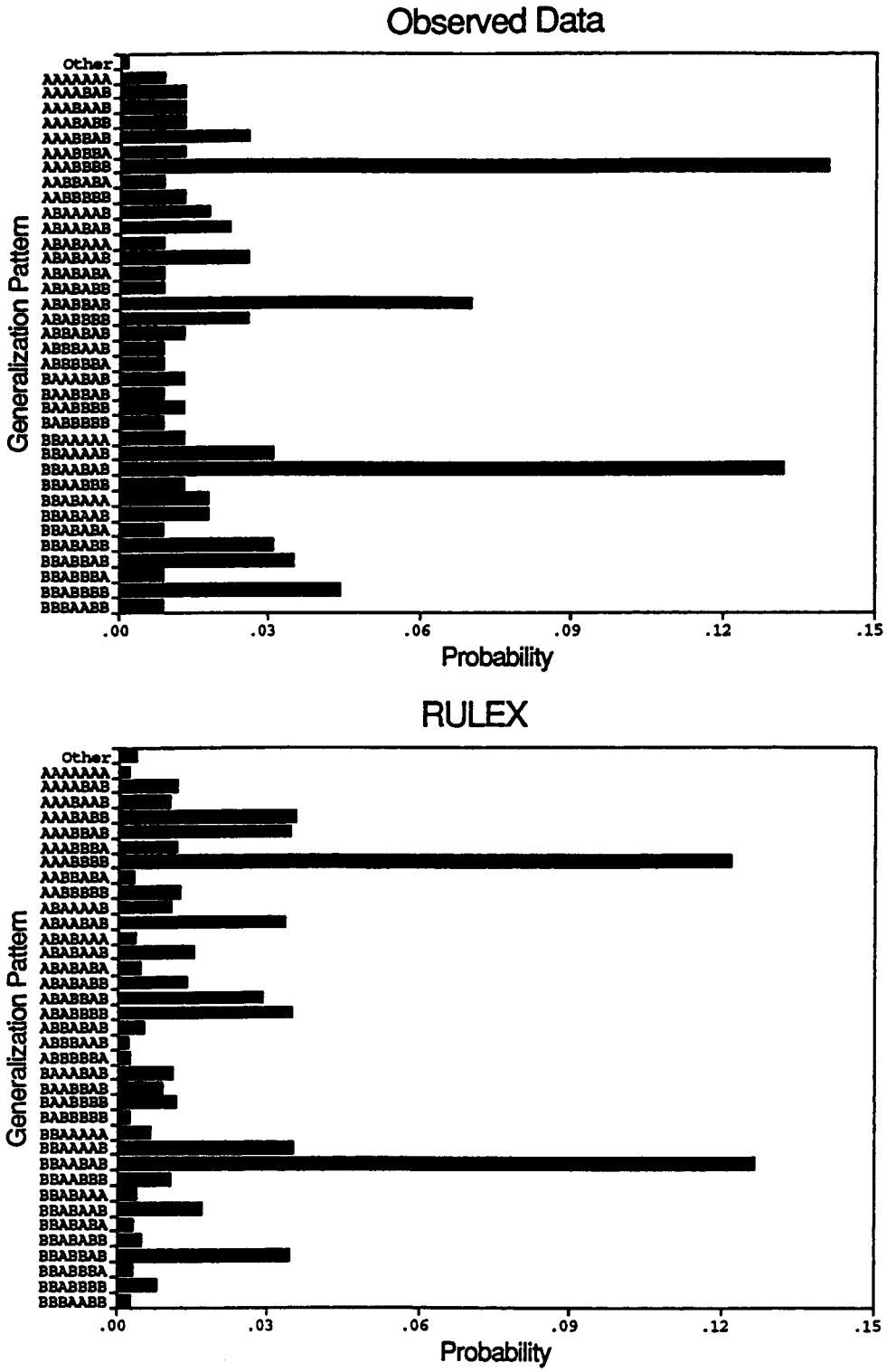

Figure 1. Top panel: Distribution of generalization profiles observed in Nosofsky, Palmeri, and McKinley's (1994) experiment. Bottom panel: Distribution of generalizations predicted by RULEX. From "Rule-Plus-Exception Model of Classification Learning," by R. M. Nosofsky, T. J. Palmeri, and S. C. McKinley, 1994, *Psychological Review, 101*, pp. 72–73 (Figures 9–10). Copyright 1994 by the Americal Psychological Association. Adapted with permission.

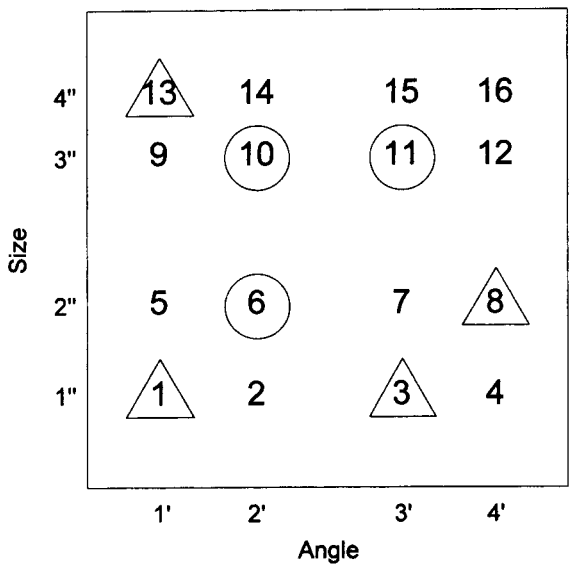

Figure 2. Category structure tested by Nosofsky, Clark, and Shin (1989). Stimuli enclosed by circles = members of Category A; stimuli enclosed by triangles = members of Category B; unenclosed stimuli = transfer items. Single primes denote values along Dimension 1; double primes denote values along Dimension 2.

sional generalization of signal detection theory. One of the key ideas is that an observer establishes decision boundaries in a multidimensional psychological space. These boundaries partition the space into response regions. Anytime a percept falls in Region A, a Category A response is made. For RULEX, which places emphasis on single-dimension rules, the decision boundaries take a highly simplified form: They are simply linear boundaries that are orthogonal to the coordinate axes of the space. (For previous examples in which continuous-dimension logical rules have been formalized in terms of orthogonal linear boundaries, see Nosofsky et al., 1989, and Nosofsky, 1991.) For example, in the Figure 2 category structure, an observer might establish a criterion on the dimension of angle, such that any object with an angle level of 2 or greater is classified into Category A and any object with an angle level of 1 or less is classified into Category B. We summarize this rule by using the notation A: $\geq 2'$, where the single prime indicates that we are referring to values along Dimension 1 (angle). Likewise, the rule A: $\geq 3''$ indicates that members of Category A are those with values greater than or equal to 3 on Dimension 2 (size). Of course, to solve the categorization problem, exceptions would need to be learned for each of these rules, a process that we discuss below.

The rules just described involve the setting of a single criterion. Another important type of single-dimension rule occurs when the observer establishes an interval defined by two decision criteria. Any percept falling within the interval is classified in one category, and any percept falling outside the interval is classified in the alternative category. For example, a likely double-criterion rule for the Figure 2 structure would be B: $\leq 1' \vee \geq 4'$ (members of Category B are those with values of ≤ 1 OR ≥ 4 on Dimension 1). For simplicity, in developing the continuous-dimension version of RULEX, we limit initial consideration to these two types of single-dimension rules (i.e., single criterion and double criterion). Later, we consider more complex logical rules that result from combinations of orthogonal linear boundaries along multiple dimensions, such as conjunctive, disjunctive, and biconditional rules.

Because of perceptual noise in the object representation or criterial noise in the location of the decision boundary, classification of items into categories according to the rules may not be completely deterministic. Suppose that a given Rule K is used. Then the probability that item i is classified into Category A is denoted $P_{\text{Rule K}}(\text{A}|i)$ and is found by integrating over the portion of the item i distribution that falls in the Response A region defined by Rule K. In the current model, a single perceptual-criterial noise parameter, σ, is assumed when one is estimating these probabilities.[1]

Once the single-dimension rule is established, the observer needs to store exceptions, just as occurs in the discrete, binary-valued dimension case. When continuous-dimension stimuli are involved, however, it becomes critical to consider the role of stimulus similarity in guiding the *use* of stored exceptions. In the binary-valued version of RULEX, we assumed, for simplicity, that a stored exception was used to classify an object only if it perfectly matched the object on its relevant attributes. For example, learning the exception 2*11 → A would lead the observer to classify only objects 2111 and 2211 into Category A. For continuous-dimension stimuli, we propose that a similarity-comparison process is used to guide classifications based on exceptions. For example, if the observer learns that an object with size 5 mm and angle 46° is an exception that belongs in Category A, it is likely that the observer would also classify an object with size 5.1 mm and angle 46.3° into Category A.

Suppose that the observer has learned Rule K for partitioning the space, and let E(K) denote the set of exceptions that a learner has stored in memory to completely solve the problem (given this rule). For example, if the observer adopted the rule A: $\geq 2'$, then he or she would need to learn that Stimuli 3 and 8 are exceptions that belong to B (see Figure 2). (In the examples considered in this article, the stimuli are composed of two dimensions, so storing an exception to the rule amounts to storing a complete exemplar. However, just as is the case in the binary-valued version of RULEX, when multiple dimensions compose the objects, the exceptions that are stored could consist of subsets of dimensions of the complete exemplars.) Following previous work, we assume that the probability that the exception process is used to classify item i is related to the summed similarity of item i to all exceptions k belonging to E(K). Specifically, the probability that the exception process is used to classify item i is given by

$$P_{\text{E use}}(\text{K}) = \frac{\left[\sum_{k \in \text{E(K)}} s(i,k)\right]}{\left\{\left[\sum_{k \in \text{E(K)}} s(i,k)\right] + v\right\}}, \quad (1)$$

where $s(i,k)$ denotes the similarity of item i to exception k, and v represents a criterion for the use of the exception information (cf. Estes & Maddox, 1995; Nosofsky, 1988).

Similarity is computed in the model by using classic methods from multidimensional scaling (MDS) theory (Shepard, 1958, 1964, 1987; see Nosofsky, 1992, for a review of the use of these methods in modern exemplar models). The stimuli represented in Figure 2 were readily discriminable and varied along highly separable dimensions. For such stimuli, we assume that the similarity between item i and exception k is an exponential decay function of their distance in the multidimensional space (Shepard, 1987):

$$s(i,k) = \exp[-\kappa \cdot d(i,k)], \quad (2A)$$

where κ is a freely estimated scaling parameter. Furthermore, for these highly separable-dimension stimuli, we assume that distance $d(i,k)$ is computed by using a city-block metric (see, e.g., Garner, 1974; Nosofsky et al., 1989; Shepard, 1991):

$$d(i,k) = \sum_m |x_{im} - x_{km}|, \quad (2B)$$

where x_{im} denotes the psychologically scaled value of item i on dimension m.

Table 2
Average Probability With Which Each Stimulus Was Classified in Category A in Nosofsky, Clark, and Shin's (1989) Experiment 1, Together With the Predictions From RULEX

Stimulus	Obs	Pre-1	Pre-2
B1	.05	.06	.06
2	.54	.56	.56
B3	.17	.16	.18
4	.06	.05	.05
5	.19	.22	.20
A6	.83	.84	.82
7	.37	.41	.42
B8	.08	.08	.09
9	.18	.19	.19
A10	.86	.90	.90
A11	.91	.86	.85
12	.39	.38	.40
B13	.08	.08	.08
14	.76	.80	.81
15	.87	.85	.83
16	.32	.36	.36

Note—Obs, observed probabilities. Pre-1, predicted probabilities from RULEX when the model is fitted to the averaged classification transfer data by using the AIC statistic. Pre-2, predicted probabilities from RULEX when the model is fitted simultaneously to the averaged classification transfer data and the distribution-of-generalizations data by using the composite *SSD*(C) measure. Training exemplars from Categories A and B are denoted with an A or a B, respectively.

Once the exception process is invoked, the probability that item i is classified in Category A is given by

$$\text{P}_{\text{exc}}(\text{A} \mid i) = \frac{\sum_{k \in \text{EA(K)}} s(i,k)}{\sum_{k \in \text{EA(K)}} s(i,k) + \sum_{k \in \text{EB(K)}} s(i,k)}, \quad (3)$$

where EA(K) and EB(K) denote the set of exceptions belonging to Categories A and B, respectively.

Bringing these ideas together, suppose that the observer has learned Rule K and has stored the set of exceptions corresponding to that rule, E(K). With probability $P_{\text{E use}}(\text{K})$, item i invokes the exception-use process; and with probability $[1 - P_{\text{E use}}(\text{K})]$, the rule is used. Thus, the probability that the observer classifies item i into Category A is given by

$$P_{\text{K}}(\text{A} \mid i) = P_{\text{E use}}(\text{K}) \cdot P_{\text{exc}}(\text{A} \mid i) + \left[1 - P_{\text{E use}}(\text{K})\right] \cdot P_{\text{Rule K}}(\text{A} \mid i). \quad (4)$$

Finally, because multiple sets of rules and exceptions are available for solving any given problem, the averaged classification data are predicted by summing over the probabilities that each individual Rule K is formed. Thus, letting $P(\text{K})$ denote the probability that any given observer forms Rule K, the overall probability that a group of observers classifies item i into Category A is given by

$$P(\text{A} \mid i) = \sum_{\text{K}} [P(\text{K}) \cdot \text{P}_{\text{K}}(\text{A} \mid i)]. \quad (5)$$

The free parameters in this version of the RULEX model are the set of rule probabilities, $P(\text{K})$; the noise involved in applying the rules (σ); the similarity-scaling parameter κ; and the exception-use criterion v. As will be seen in the applications of the model, a reasonably small number of candidate rules can often be hypothesized, so the model is testable. The precise locations of the decision boundaries in the multidimensional space can also be treated as free parameters, but, at least for the initial problems considered in this article, these locations can be set at reasonable default values with essentially no effect on the overall model fits.

Application to Nosofsky, Clark, and Shin (1989)

Our first test of RULEX is obtained by fitting the model to the averaged classification data obtained by Nosofsky et al. (1989, Experiment 1) for the category structure in Figure 2. In Nosofsky et al.'s (1989) experiment, there was an initial training phase in which only the seven assigned category exemplars were presented. Corrective feedback was provided on every trial. Following training, there was a test phase in which all 16 stimuli in the set were presented five times each with no feedback. The probability with which each stimulus was classified into Category A is given in Table 2. The transfer data are only for the observers who achieved a reasonably strict learning criterion during the initial training phase. Thus, we assume that most or all of these observers had learned rules and exceptions that enabled them to solve the problem.

Table 3
List of Rule-Plus-Exception Strategies for the Nosofsky, Clark, and Shin (1989) Experiment 1 Category Structure

#	Rule	Exceptions
1	B: $\leq 1' \vee \geq 4'$	3
2	B: $\leq 1'' \vee \geq 4''$	8
3	B: $\leq 2''$	6, 13
4	B: $\leq 1'$	3, 8
5	B: $\leq 1''$	8, 13
6	B: $\leq 1' \vee \geq 3'$	11
7	B: $\leq 2'' \vee \geq 4''$	6

To apply RULEX, we assume that only the single-dimension rules that give rise to no more than two exceptions have nonzero probabilities. This assumption is akin to the criterion used in the binary-valued version of RULEX, in which rules were retained only if they yielded reasonable levels of performance (Nosofsky, Palmeri, & McKinley, 1994, pp. 55–58). There are seven such rules available, and they are listed in Table 3 along with the set of exceptions necessary for each rule. Because the rule probabilities are constrained to sum to 1.0, there are six free rule-probability parameters, plus the parameters σ, κ, and v. In all cases, the locations of the rule boundaries were set midway between the adjacent dimension values where they were positioned.

We fitted RULEX to the transfer data by searching for the free parameters that minimized Akaike's information criterion (AIC) statistic (Akaike, 1974), given by

$$\text{AIC} = -2 \ln L + 2N, \quad (6)$$

where $\ln L$ is the (natural) log likelihood of the data given the model, and N is the number of free parameters in the model.[2] Smaller values of the AIC statistic reflect a better fit for a model. Note that the AIC statistic penalizes a model for the number of free parameters that it uses. Although sole reliance on the AIC statistic has certain pitfalls, we used it as a preliminary guide to help evaluate the fits of competing models with different numbers of free parameters.[3]

The predicted probabilities from RULEX are shown alongside the observed probabilities in Table 2. The model yielded AIC = 92.0, accounted for 99.4% of the variance in the classification response probabilities, and achieved a root-mean-squared deviation (*RMSD*) with the observed classification probabilities of .026.

As a source of comparison, we also fitted Nosofsky's (1986) *generalized context model* (GCM) to the classification data. The GCM generalizes Medin and Schaffer's (1978) exemplar-based context model to the domain of continuous-dimension stimuli. In the present case, it uses three free parameters to fit the data: a similarity-scaling parameter, κ, an attention weight, w_1, and a response-bias parameter, β_1 (see Nosofsky et al., 1989, for a detailed discussion of the model as applied to the present data). The GCM yielded AIC = 134.4 and accounted for 97.8% of the variance in the response probabilities (*RMSD* = .048). By conventional criteria, the fit of the GCM would be considered excellent. Nevertheless, the superior fit of the continuous-dimension RULEX model sets a new standard. Although RULEX uses substantially more free parameters than does the GCM in the present situation, the superior AIC fit achieved by RULEX provides some initial clues that these extra free parameters may be doing some important work.

More impressive than its ability to fit the averaged classification transfer data is that RULEX also does an excellent job of predicting the distribution of individual-observer generalizations. The observed distribution, computed previously by Nosofsky et al. (1989, p. 291), is reported in Table 4. This distribution is based on 122 subjects who satisfied the learning criterion established in the training phase. For example, 21 of the 122 learners (17.2%) displayed generalization profile ABBABBAAB, meaning that, during the test phase, they classified Transfer Stimuli 2, 7, 14, and 15 in Category A and Transfer Stimuli 4, 5, 9, 12, and 16 in Category B—see Figure 2. (Following the spatial layout in Figure 2, the transfer stimuli in each profile are listed in the order 2, 4, 5, 7, 9, 12, 14, 15, 16.)

The distribution of generalizations predicted by the present version of RULEX, with its best-fitting parameters held fixed from the previous analysis in which the averaged transfer data were fitted (Table 2), is also reported in Table 4. (The manner in which the distribution of generalizations is predicted from the model is explained in Appendix A.) The model accounts for 82.4% of the variance in the distribution-of-generalizations data. Thus, RULEX does a reasonably good job of simultaneously characterizing both the averaged transfer data and the distribution of behavior at the individual-observer level.

It is critical to understand that the ability of RULEX to fit the distribution of generalizations is not an automatic consequence of its excellent fit to the averaged transfer data. Indeed, a wide variety of different distributions are consistent with the same averaged transfer data. As one example, consider a model that predicts the averaged transfer data *perfectly*, and which assumes that all individual observers behave identically (except for random noise).

Table 4
Observed and Predicted Distribution of Generalization Data From Nosofsky, Clark, and Shin's (1989) Experiment 1

Profile	Observed	RULEX-1	PFM	RULEX-2
ABBABBAAB	.172	.176	.068	.178
ABBBBBAAB	.156	.265	.134	.232
BBBABBAAB	.082	.037	.053	.041
BBBBBBAAB	.074	.056	.106	.054
ABBBBAAAA	.066	.057	.030	.063
BBABAABAA	.041	.008	.000	.008
ABBBBBAAA	.033	.012	.053	.017
Other	.377	.388	.556	.407

Note—RULEX-1, RULEX's predicted probabilities derived from the use of the AIC statistic. PFM, predicted probabilities from the perfect-fitting average-probability model. RULEX-2, RULEX's predicted probabilities derived from the use of the composite *SSD*(C) measure.

Table 5
Best-Fitting Parameters Obtained by Fitting RULEX Simultaneously to the Averaged Transfer Data and Distribution-of-Generalization Data in Nosofsky, Clark, and Shin (1989) and in Experiments 1 and 2

	Data Set		
Parameter	1	2	3
R1	.51	.32	.00
R2	.00	.00	.50
R3	.17	.14	.00
R4	.13	.38	.03
R5	.13	.00	.46
R6	.00	.15	.00
R7	.07	.01	.00
v	.31	.17	.17
κ	.37	2.44	3.34
σ	.00	.00	.00

Note—Data set: 1, Nosofsky, Clark, and Shin (1989); 2, Experiment 1; 3, Experiment 2. Parameters R1–R7 denote probabilities of Rules 1–7, respectively. v, criterion for use of exception information; κ, similarity-scaling parameter; σ, perceptual-criterial noise parameter.

In other words, all individual observers have the same probability vector for the 16 stimuli, and it is identical to the averaged probability vector. The distribution of generalizations predicted by this model is given in the third column in Table 4. This model accounts for only 46.2% of the variance in the observed distribution-of-generalization data.

The key lesson here is that, according to RULEX, individual observers may differ greatly in the classification rules that they adopt, especially when multiple rules are available for solving the problem. Averaged classification data often represent a mixture of an array of idiosyncratic rules and exceptions; they do not arise from a homogeneous distribution at the individual-observer level. Furthermore, these analyses suggest that RULEX does quite well at characterizing this range of individual-observer strategies giving rise to the averaged data. (For previous examples from the classification literature that place emphasis on the importance of modeling the heterogeneity in individual-observer performance, see Ashby, Maddox, & Lee, 1994; Martin & Caramazza, 1980; Nosofsky et al., 1989; Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995; J. D. Smith, Murray, & Minda, 1997.)[4]

For completeness, we also conducted an analysis in which RULEX was fitted simultaneously to the averaged transfer data in Table 2 and the distribution-of-generalization data in Table 4. Such an analysis is important because multiple parameter settings may be available that can fit the averaged transfer data. Therefore, holding these parameters fixed may, in some situations, greatly underestimate the ability of RULEX to describe the distribution of generalizations. Any method for combining the fits to the averaged transfer data and the distribution of generalizations into a composite measure is arbitrary. After preliminary exploration, we defined *SSD*(T) as the sum-of-squared deviations between the predicted and observed classification transfer probabilities, *SSD*(D) as the sum-of-squared deviations between the predicted and observed probabilities for the distribution of generalizations, and *SSD*(C) as a composite measure given by $SSD(C) = SSD(T) + 4 \cdot SSD(D)$. We then searched for the free parameters that minimized *SSD*(C). The results are reported in the final columns of Tables 2 and 4, which show RULEX's predicted probabilities for each data set. RULEX yielded *SSD*(C) = .069 and accounted for 99.2% of the variance in the averaged classification transfer probabilities [*SSD*(T) = .028] and for 88.6% of the variance in the distribution-of-generalizations data [*SSD*(D) = .010]. We consider these excellent fits to provide support for the model.

The best-fitting parameters derived from the fit of RULEX to the composite data are reported in Table 5.

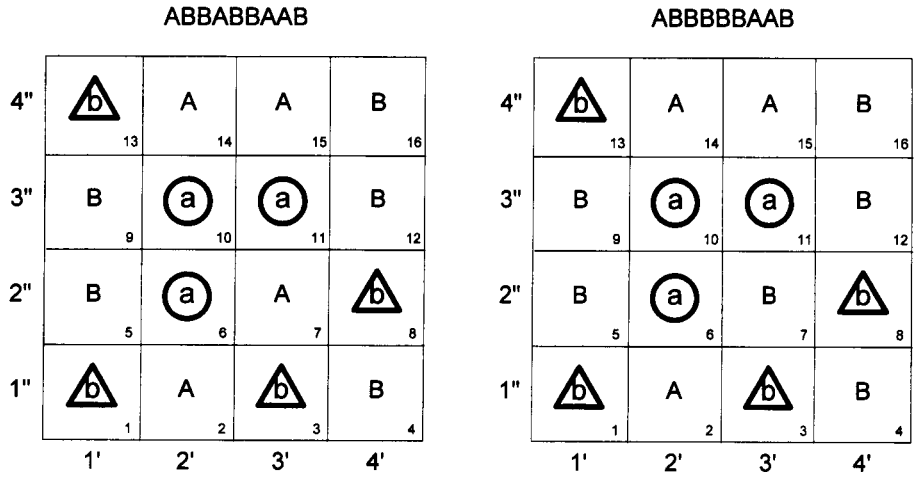

Figure 3. Generalization profiles most likely to emerge from the use of Rule 1 (see Table 3). Single primes denote values along Dimension 1; double primes denote values along Dimension 2.

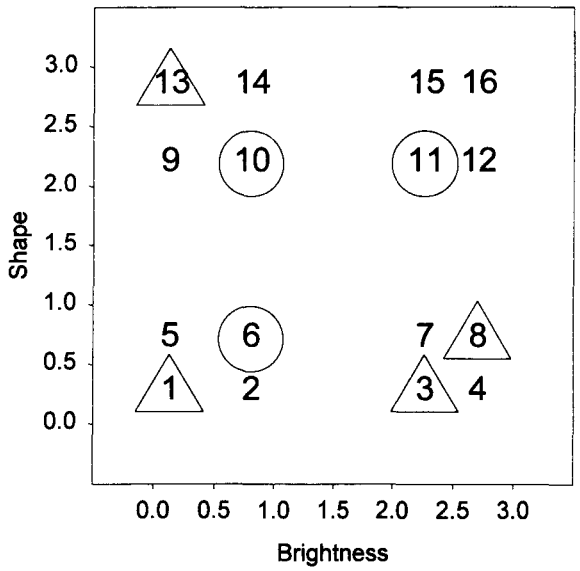

Figure 4. Constrained two-dimensional scaling solution for the ellipses used in Experiment 1.

One point of interest is that several of the free parameters had best-fitting values of zero, so in essence the model is making use of fewer "effective" free parameters than listed previously. For example, the best-fitting value of the perceptual-criterial noise parameter (σ) was zero, suggesting that there was very little noise involved in applying the logical rules. Recall that the stimuli used in Nosofsky et al.'s (1989) experiment were highly discriminable. We expect that the σ parameter would take on far greater importance in experimental situations involving perceptually confusable stimuli.

The parameter estimates also reveal a high probability for the use of Rule 1. Thus, according to this analysis, the most prevalent strategy adopted by the observers was to use the rule B: $\leq 1'$ V $\geq 4'$ while remembering that Training Stimulus B3 was an exception to this rule. The generalization profiles most likely to emerge from the use of this rule are illustrated in Figure 3, and they correspond precisely to the two highest frequency generalization profiles reported in Table 4.

One reason that this rule may have been prevalent involves the physical dimension values used by Nosofsky et al. (1989) in their design. Angle value 1′ was a nearly horizontal line pointing to the right, whereas angle value 4′ was a nearly horizontal line pointing to the left. The intermediate angle values (2′ and 3′) pointed in a more upwards direction. Thus, from a psychological perspective, the rule B: 1′ V 4′ can be summarized by the simpler rule: "Respond Category B if the angle is nearly horizontal." Perhaps the availability of this extremely simple verbal rule promoted the observers' use of the RULEX strategy. To assess the generality with which the model may apply, we decided to repeat the Nosofsky et al. (1989) experiment, but using a new stimulus set in which this type of very simple rule was unavailable.

EXPERIMENT 1

In this experiment, we again used the category structure illustrated in Figure 2, except that instead of using circles varying in size and angle of radial line, we used as stimuli a set of ellipses varying in their shape and brightness. Extensive similarity-scaling work was first conducted to find physical dimension values that yielded a psychological structure close to the one shown in Figure 2. We then conducted the classification learning and transfer test on a separate group of subjects. The goal was to use RULEX to once again fit the averaged transfer data and the distribution of generalizations.

Method

Subjects. A group of 30 subjects was tested in a similarity-scaling study to verify the dimensional structure of the stimulus set that we constructed. Another 185 subjects were tested in the classification learning study. All subjects were undergraduates at Indiana University who received partial credit toward an introductory psychology course requirement.

Stimuli and Apparatus. The stimuli were 16 ellipses created by factorially combining four levels of shape and four levels of brightness. Extensive pilot work was conducted to find physical values of shape and brightness that yielded a psychological structure close to that shown in Figure 2. Shape was defined as the ratio of the width to the height of each ellipse. The four ratio values used were 5.692:1, 4.125:1, 3.000:1, and 2.190:1. The areas of the ellipses were equated, with the narrowest ellipse spanning approximately 4¼ in. by ¾ in. The brightness of the ellipses varied from black (0) to dark gray (50) to light gray (160) to white (250). The values in parentheses indicate the intensities of the red, green, and blue channels on the video board (255 maximum) on the CompuAdd 486 personal computers. The ellipses were displayed against a bright red background on 14-in. computer monitors.

Procedure. In the similarity-scaling study, all 120 distinct pairs of the 16 ellipses were presented four times each during the course of 480 trials. On each trial, 2 distinct ellipses were presented side by side on the screen, and the subject judged their similarity on a scale from 1 (*most dissimilar*) to 9 (*most similar*). The subjects were urged to use the full range of ratings in making their judgments. The ellipse pairs were presented in four blocks of 120 trials each, with each unique pair presented once per block in a random order.

The classification-learning experiment used a repeating training–test procedure. The full sequence consisted of 10 blocks of training trials, 3 blocks of test trials, 10 blocks of training trials, 3 blocks of test trials, 20 blocks of training trials, and 3 blocks of test trials. During each training block, each of the seven assigned category exemplars was presented once in a random order. On each trial, the subject classified the ellipse into either Category A or Category B by pressing a button on the computer keyboard, and corrective feedback was then presented on the screen. During each test block, all 16 stimuli were presented in a random order. The subjects again classified each ellipse into Category A or Category B, but no corrective feedback was provided.

Results and Theoretical Analysis

Similarity scaling. A constrained two-dimensional solution for the ellipses was derived from the matrix of

Table 6
Predicted and Observed Classification Transfer Data From the Final Set of Test Blocks in Experiment 1

	Category A Response Probability	
Stimulus	Pre	Obs
B1	.042	.010
2	.843	.866
B3	.113	.057
4	.083	.016
5	.080	.037
A6	.922	.943
7	.227	.171
B8	.067	.071
9	.082	.094
A10	.975	.978
A11	.924	.917
12	.576	.539
B13	.022	.029
14	.915	.947
15	.905	.917
16	.552	.539

Note—Pre, predicted; Obs, observed. Training stimuli from Categories A and B are denoted with an A or a B, respectively.

averaged similarity ratings. In this solution, all ellipses with a common physical value of brightness were constrained to have the same psychological coordinate on the brightness dimension, and all ellipses with the same physical shape were constrained to have the same psychological coordinate on the shape dimension. A city-block metric was used for computing distance in the space. The constrained two-dimensional solution makes use of only six free parameters for fitting the matrix of 120 similarity judgments. The MDS model was fitted to the data by searching for the psychological coordinates that minimized stress (see, e.g., Kruskal & Wish, 1978). The MDS solution yielded a stress of only .060, which is considered quite a good fit even for standard MDS solutions that allow coordinate parameters for all stimuli to vary freely (without constraints imposed by the physical structure of the stimulus set). The constrained two-dimensional solution is illustrated graphically in Figure 4. The structure of the derived space corresponds closely to the planned design.

Classification. Because the continuous-dimension version of RULEX is applicable only in situations in which observers have formed rules and exceptions that solve accurately the classification problem, we established a learning criterion for including each subject's data in the modeling analyses. Specifically, we eliminated from analysis any subject who made greater than 15% errors during the final 10 training blocks. Use of this criterion led to the removal of 11.4% of the subjects, leaving a total of 164 subjects.

The probability with which the learners classified each of the 16 stimuli into Category A during the final three test blocks is reported in Table 6. The distribution of generalizations for the learners is reported in Table 7.[5]

In an initial analysis, we fitted RULEX and the GCM to the classification transfer data by searching for the parameters that minimized the AIC statistic. RULEX yielded AIC = 86.6, accounting for 99.8% of the variance in the response probabilities (*RMSD* = .018). The GCM yielded AIC = 109.28, accounting for 99.5% of the variance (*RMSD* = .030). RULEX again yields a better AIC fit to the classification transfer data than does the GCM.

Next, we fitted RULEX simultaneously to the classification transfer data and distribution-of-generalizations data by using the composite measure of fit described previously. The predicted probabilities are shown alongside the observed probabilities in Tables 6 and 7. RULEX yielded *SSD*(C) = .059, accounting for 99.3% of the variance in the classification transfer data (*RMSD* = .033), and 94.6% of the variance in the distribution-of-generalizations data (*RMSD* = .024). Once again, we interpret these excellent fits as providing support for the proposed RULEX model.

The best-fitting parameters for RULEX are reported in Table 5. In the Nosofsky et al. (1989) experiment, the use of Rule 1 dominated the observers' behavior, whereas in the present experiment, there was a greater mix of alternative RULEX strategies. This result supports our hypothesis that recoding of the dimension values to yield a rule based on horizontal angles may have occurred in the Nosofsky et al. (1989) study. In the present experiment, the most prevalent rules were Rules 1, 3, 4, and 6 (see Tables 3 and 5). To gain some insight into these results, note from Table 7 that the most common generalization profiles were ABBBBAAAA and ABBBBBAAB. We illustrate these profiles in Figure 5. Profile ABBBBAAAA is the highest probability profile predicted by Rule 4, and profile ABBBBBAAB is the highest probability profile predicted by Rule 1.

Inspection of the rule-probability parameter estimates also reveals that both single-criterion and double-criterion rules were adopted by the observers (see Tables 3 and 5). Although single-criterion rules may be less cognitively complex than double-criterion rules, note that in the present design the double-criterion rules required the storage of only one exception, whereas the single-criterion rules required the storage of two exceptions. These factors prob-

Table 7
Predicted and Observed Distribution-of-Generalization Data From the Final Set of Test Blocks in Experiment 1

	Probability	
Profile	Pre	Obs
BBBBBBAAB	.013	.018
BBBBBAAAA	.023	.024
ABBBBBAAB	.235	.201
ABBBBBAAA	.022	.067
ABBBBAABB	.036	.012
ABBBBAAAB	.047	.012
ABBBBAAAA	.291	.287
ABBBAAAAA	.010	.018
ABBABBAAB	.081	.091
Other	.243	.268

Note—Pre, predicted; Obs, observed.

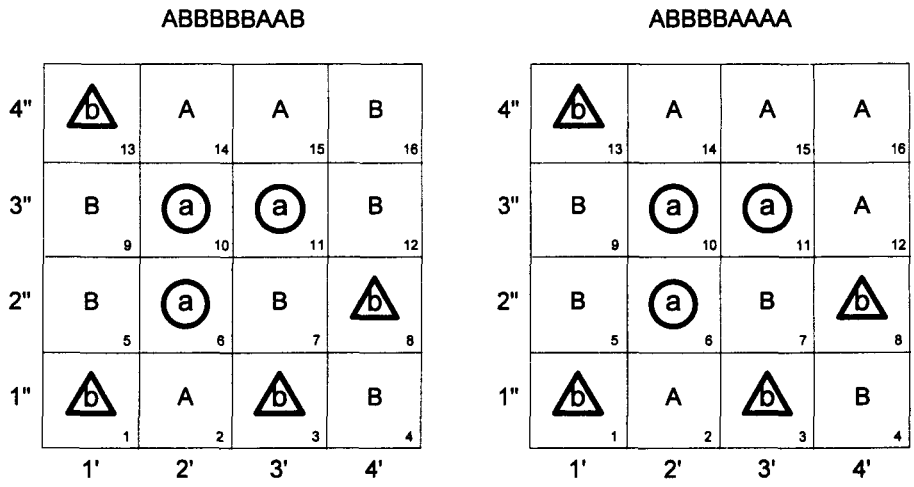

Figure 5. The most common generalization profiles observed in Experiment 1. Profile ABBBBAAAA is the highest probability profile predicted by Rule 4, and profile ABBBBBAAB is the highest probability profile predicted by Rule 1. Single primes denote values along Dimension 1; double primes denote values along Dimension 2.

ably trade off in influencing the types of RULEX strategies that observers adopt.

The probability with which the learners classified each of the 16 stimuli into Category A during the early sets of test blocks is reported in Table 8. Because we have not developed a learning version of the continuous-dimension RULEX model, we cannot formally model these data. However, the pattern of early transfer data provides some converging evidence for the type of learning process that we envision. Recall that our modeling analyses indicated that the most prevalent rules adopted by the observers were Rules 1 and 4, which occurred with high frequency, and Rules 3 and 6, which occurred with lower frequency. As indicated in Table 3, Training Stimulus B3 is the only exception to Rule 1, and it is one of the two exceptions to Rule 4. Interestingly, as revealed by the early test data in Table 8, Stimulus B3 had the most errors among the Category B training stimuli. By contrast, Training Stimulus B1, which was not an exception to any of the rules, had the fewest errors among the Category B training stimuli. Likewise, Training Stimuli A6 and A11 are exceptions to Rules 3 and 6, respectively, whereas Training Stimulus A10 is not an exception to any of the rules. Interestingly, Training Stimulus A10 had the fewest errors among the Category A training stimuli. These patterns of results are consistent with the idea that many of the observers had developed hypotheses based on Rules 1, 4, 3, and 6 by the early test blocks, but had not yet formed the exceptions necessary to solve the classification problem. We also observed that the distributions of generalizations during the early test blocks were quite a bit more diffuse than the distribution observed in the asymptotic data. Such a pattern is expected because, early in learning, the observers are still exploring a multitude of different rules and are at different stages of exception storage as well.

Although in comparison with Nosofsky et al.'s (1989) results, there was a reduction in the use of Rule 1 in the present experiment, note that the most prevalent rules still tended to be defined along Dimension 1 (see Tables 3 and 5). This result led us to wonder whether the focus on Dimension 1 had something to do with the abstract category structure or with the particular physical dimensions that were used to instantiate the abstract structure. For example, perhaps the brightness dimension is more "salient" than the shape dimension, or perhaps it is easier to verbalize rules based on the present brightness values than on the present values of ellipse shape. To investigate this issue, we again tested the Figure 4 structure,

Table 8
Observed Category A Response Probabilities From the First Two Test Blocks of Experiment 1

	Test Block	
Stimulus	1	2
B1	.067	.035
2	.709	.823
B3	.197	.116
4	.065	.043
5	.079	.059
A6	.783	.880
7	.279	.232
B8	.136	.077
9	.183	.130
A10	.898	.953
A11	.746	.852
12	.533	.526
B13	.134	.063
14	.850	.892
15	.839	.884
16	.563	.549

Note—Training stimuli from Categories A and B are denoted with an A or a B, respectively.

Table 9
Predicted and Observed Classification Transfer Data From the Final Set of Test Blocks in Experiment 2

	Category A Response Probability	
Stimulus	Pre	Obs
B1	.000	.008
2	.038	.011
B3	.006	.014
4	.015	.008
5	.954	.903
A6	.994	.964
7	.587	.575
B8	.141	.083
9	.728	.750
A10	.915	.953
A11	.992	.981
12	.935	.908
B13	.066	.028
14	.236	.286
15	.496	.486
16	.491	.478

Note—Pre, predicted; Obs, observed. Training stimuli from Categories A and B are denoted with an A or a B, respectively.

except that we switched the assignment of physical dimensions to the logical dimensions.

EXPERIMENT 2

The procedure in Experiment 2 was similar to that in Experiment 1; the main difference being that in the present experiment, Dimension 1 corresponded to ellipse shape, whereas Dimension 2 corresponded to brightness. If something about the logical structure of the categories led subjects to focus on Dimension 1 in the previous experiment, then the same pattern of generalization and rule-use estimates should be obtained in this experiment. On the other hand, new patterns of generalization should be obtained if observers preferred to form rules based on brightness than on ellipse shape.

Method

Subjects. The subjects were a new set of 185 undergraduates from Indiana University who received partial credit toward an introductory psychology course requirement.

Stimuli and Apparatus. The stimuli and apparatus were the same as those used in Experiment 1.

Procedure. The procedure was the same as in Experiment 1, with the following exceptions. First, we switched the assignment of the physical dimensions of brightness and shape to the abstract category structure. Dimension 1 corresponded to shape, and Dimension 2 corresponded to brightness. Second, because overall performance on the training stimuli was extremely high in Experiment 1, we were worried that the experiment might be tedious, so we reduced the number of training and transfer blocks. The complete sequence consisted of 5 training blocks, 1 transfer block, 5 training blocks, 1 transfer block, 10 training blocks, and 3 transfer blocks.

Results

We eliminated from analysis any subject who made more than three errors during the final five training blocks. Use of this criterion led to the removal of 35% of the subjects, leaving a total of 120 subjects. Our ensuing conclusions do not change if we adopt a more lax criterion and fit RULEX to the data of a larger proportion of observers, but we think that it is important to restrict the modeling analyses to the set of observers who successfully solved the problem. The data from the nonlearners are presented and discussed in Appendix B.

The probability with which the learners classified each of the stimuli into Category A during the final three transfer blocks is reported in Table 9, and the distribution-of-generalization data are reported in Table 10. Upon inspection of the data, it is immediately apparent that switching the assignment of physical dimensions had a dramatic effect on the pattern of results. For example, Transfer Stimuli 5 and 9 had a strong tendency to be classified in Category B in Experiment 1, whereas the reverse was observed in the present experiment. Likewise, the classification patterns for Transfer Stimuli 2 and 14 reversed dramatically across the two experiments. In addition, the structure of the distribution of generalizations was dramatically altered. As will be seen, a good explanation of these results is that observers had a strong preference for developing rules along the physical dimension of brightness rather than the physical dimension of ellipse shape.

We fitted RULEX simultaneously to the averaged transfer data and the distribution-of-generalization data by using our composite measure. The predicted probabilities are shown alongside the observed probabilities in Tables 9 and 10. RULEX yielded $SSD(C) = .108$, accounting for 99.4% of the variance in the classification transfer data ($RMSD = .031$) and 72.3% of the variance in the distribution-of-generalizations data ($RMSD = .039$). Although the fit to the distribution-of-generalization data is not quite as good as in the previous experiments, it still seems quite respectable.

The best-fitting parameters are reported in Table 5. The results indicate that by far the most prevalent strategies were to use Rules 2 and 5. Both of these rules are defined along Dimension 2 (see Table 3), which, in the pres-

Table 10
Predicted and Observed Distribution-of-Generalization Data From the Final Set of Test Blocks in Experiment 2

	Probability	
Profile	Pre	Obs
BBBABAAAA	.000	.017
BBBAAABAA	.000	.017
BBABBAAAA	.032	.033
BBABAABBB	.179	.175
BBABAABAA	.050	.083
BBABAAAAA	.032	.017
BBAABAAAA	.057	.133
BBAAABBBB	.003	.050
BBAAAABBB	.314	.267
BBAAAABAA	.088	.075
BBAAAAAAA	.056	.017
ABBBBAAAA	.013	.017
Other	.175	.100

Note—Pre, predicted; Obs, observed.

Table 11
Observed Category A Response Probabilities From the First Two Test Blocks of Experiment 2

	Test Block	
Stimulus	1	2
B1	.092	.033
2	.108	.050
B3	.150	.033
4	.083	.033
5	.625	.808
A6	.633	.842
7	.600	.533
B8	.450	.158
9	.658	.692
A10	.750	.875
A11	.792	.967
12	.758	.850
B13	.192	.075
14	.325	.300
15	.500	.467
16	.533	.475

Note—Training stimuli from Categories A and B are denoted with an A or a B, respectively.

ent experiment, was the brightness dimension. Thus, it is apparent that, in developing a learning version of the RULEX model that predicts which rules will be adopted, we will need to incorporate factors pertaining to the intrinsic physical dimensions as well as the abstract category structure.

The probability with which the learners classified the stimuli into Category A during the early test blocks is reported in Table 11. Because Rules 2 and 5 were dominant in this experiment and Training Exemplar B8 is the single exception to Rule 2 and is one of the two exceptions to Rule 5, the prediction is that the observers should have the greatest tendency to misclassify Training Exemplar B8 during the early test blocks. This prediction is strongly supported by the early test data (see Table 11). The table also reveals fairly high error probabilities for Training Exemplars A6 and B13. Note that Training Exemplar A6 has the same value along Dimension 2 as does the badly misclassified Training Exemplar B8. To the extent that, during the learning process, observers were switching the location of the Dimension 2 rule boundaries to accommodate Exemplar B8, it would result in errors for Exemplar A6. Finally, Exemplar B13 is the second exception to Rule 5, so its high error rate among the Category B exemplars is consistent with the RULEX ideas as well.

In summary, these results provide additional evidence that observers may often use RULEX strategies when learning classifications and that properties intrinsic to the physical dimensions that compose the stimuli may exert a powerful influence on which rules are formed.

APPLICATION TO NOSOFSKY (1986)

Another set of data that provides a good test of the continuous-dimension RULEX model comes from a series of classification conditions reported by Nosofsky (1986). In that experiment, the stimuli were a set of circles with an embedded radial line. The circles varied in size and in angle of the line. There were four levels of size and four levels of angle, combined orthogonally to yield a 16-member stimulus set. Unlike the type of stimuli used in Nosofsky et al.'s (1989) experiment, the stimuli were perceptually confusable.

Nosofsky (1986) tested 2 highly experienced observers across four separate classification conditions. The category structures are illustrated in Figure 6. In this figure, items enclosed by triangles represent training exemplars assigned to Category A, items enclosed by circles represent training exemplars assigned to Category B, and unenclosed items represent unassigned transfer stimuli. Following an initial learning phase in which only assigned exemplars were presented, the subjects completed extensive transfer phases in which all 16 stimuli were presented. During the transfer phase, an average of roughly 225 response observations was obtained for each individual stimulus in each categorization condition for each of the 2 observers. (See Nosofsky, 1986, pp. 43–44, for further details regarding the experimental procedure.)

The central theme in the RULEX model is that observers construct simple orthogonal linear boundaries to partition the stimulus space into category regions, and

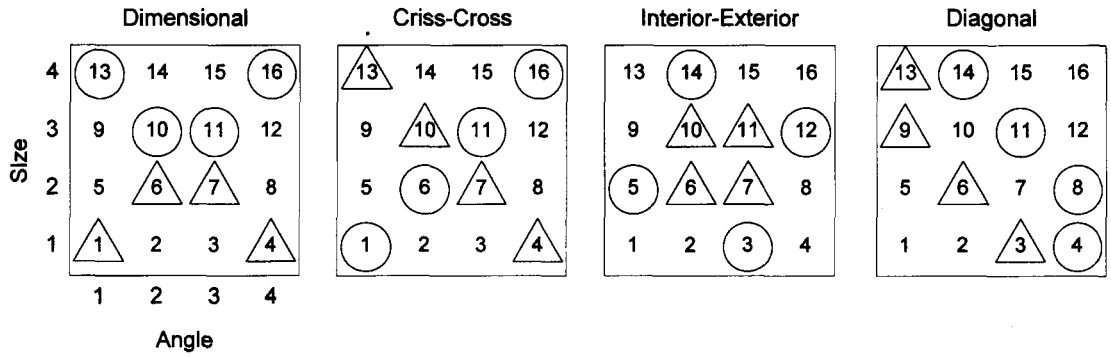

Figure 6. Category structures tested by Nosofsky (1986). Triangles, Category A training exemplars; circles, Category B training exemplars.

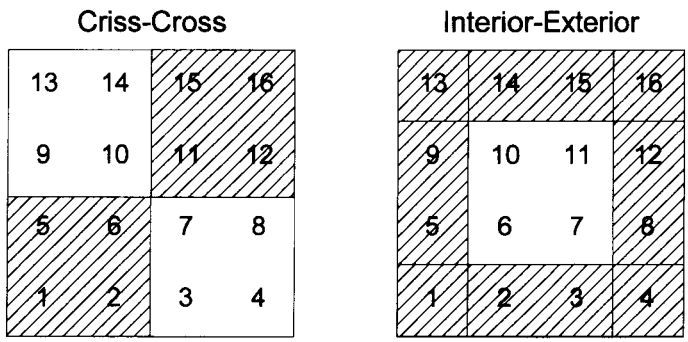

Figure 7. Left panel: Illustration of biconditional rule hypothesized for the criss-cross category structure. Right panel: Illustration of extreme-value rule hypothesized for the interior–exterior category structure.

then store exceptions to the rule. Now, in the category structures considered previously, we assumed that the orthogonal boundaries were constructed along a single stimulus dimension. In the binary-valued version of RULEX, however, single-dimension rules are used only if they work fairly well. If no single-dimension rule correctly classifies a large proportion of the stimuli, then observers are assumed to search for more complex logical rules, such as conjunctive or biconditional rules. It is straightforward to develop continuous-dimension versions of these more complex logical rules.

In applying RULEX to the category structures illustrated in Figure 6, we hypothesized that the observers used the following logical rules. For the "dimensional" categorization, the observer is assumed to place an orthogonal linear boundary along the size dimension at a location intermediate between Size Values 2 and 3. Percepts that fall above the boundary are classified in Category B, and percepts that fall below the boundary are classified in Category A. For the "diagonal" categorization, the observer is assumed to place an orthogonal linear boundary along the angle dimension at a location intermediate between Angle Values 2 and 3. The rule is to classify percepts that fall to the right of the boundary in Category B and to classify percepts that fall to the left of the boundary in Category A. In addition, the observer is assumed to store training exemplars A3 and B14 as exceptions to this rule (see Figure 6).

For the criss-cross categorization, no single-dimension rule is available. The obvious multiple-dimension rule, however, is that the observer establishes two orthogonal linear boundaries, one placed between Angle Values 2 and 3 and the other placed between Size Values 2 and 3, as illustrated in the left panel of Figure 7. Percepts falling in the shaded regions are classified in Category B and percepts falling in the unshaded regions are classified in Category A. This set of orthogonal boundaries produces a continuous-dimension biconditional rule—A: ($\leq 2'$ AND $\geq 3''$) or ($\geq 3'$ AND $\leq 2''$).

There are several plausible rule or RULEX strategies available for the interior–exterior structure, and, at the present stage of theorizing, we are unable to predict a priori which one any given observer will adopt. (Indeed, as explained in the previous sections of this article, RULEX uses a stochastic learning process for selecting among multiple rule candidates.) On the basis of preliminary model exploration, however, we posit that both observers adopted an extreme-value rule: If a percept has an extreme value on either dimension, classify it into Category B. This rule is represented by a set of four orthogonal linear boundaries, as illustrated in the right panel of Figure 7.

We fitted these RULEX models to Nosofsky's (1986) transfer data by searching for the free parameters that minimized the AIC statistic. Fitting the model to the dimensional, criss-cross, and interior–exterior categorizations required estimation of 2, 3, and 5 free parameters, respectively (i.e., the perceptual-criterial noise parameter σ, and the parameters representing the locations of the orthogonal linear boundaries needed for constructing the logical rules). Fitting the model to the diagonal categorization required estimation of five free parameters: the perceptual noise parameter σ, the location of the single orthogonal linear boundary, and three exception-use parameters.[6]

The summary fits for RULEX are given in Table 12 for each of the 2 observers in each of the four categorization conditions. For purposes of comparison, the sum-

Table 12
AIC Fits of RULEX, GCM, and QDBM to Nosofsky's (1986) Classification Transfer Data

	Model		
Condition	RULEX	GCM	QDBM
Dimensional (1)	69.2	70.6	76.7
Dimensional (2)	103.8	102.6	102.6
Diagonal (1)	123.8	132.0	110.5
Diagonal (2)	172.6	130.4	105.8
Criss-Cross (1)	146.8	208.2	112.3
Criss-Cross (2)	202.6	275.6	119.2
Interior–Exterior (1)	128.8	247.8	135.6
Interior–Exterior (2)	117.8	231.4	111.6

Note—Values in parentheses denote Participant 1 or Participant 2 in each condition. RULEX, rule-plus-exception model; GCM, generalized context model; QDBM, quadratic decision-bound model.

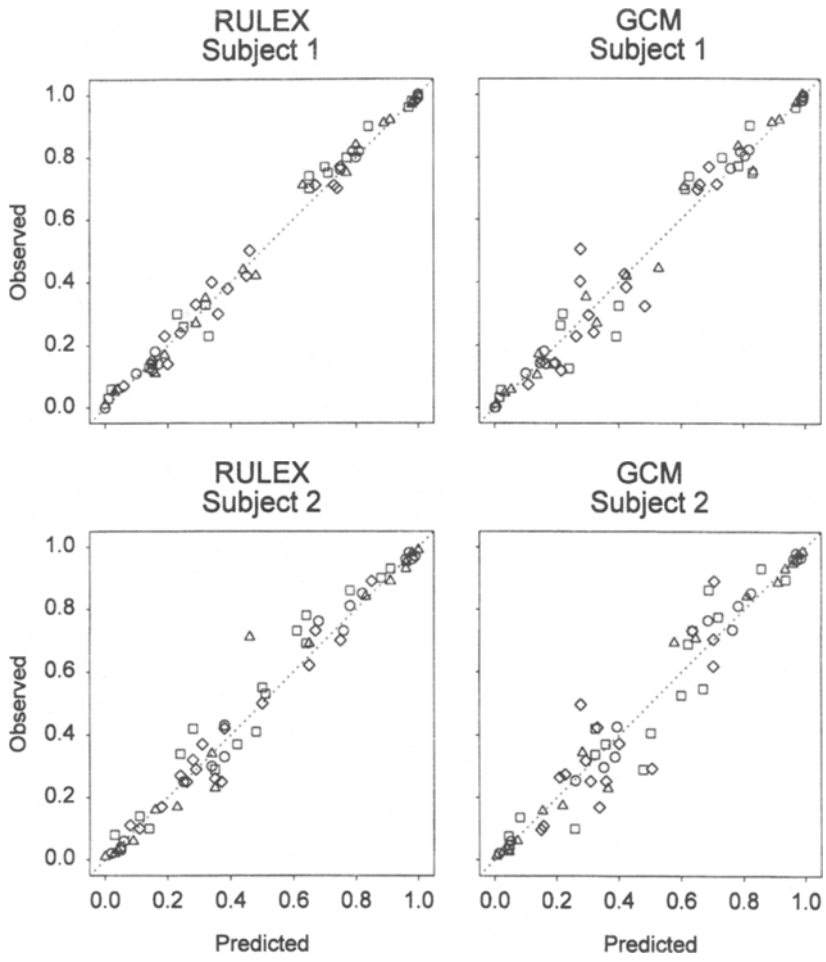

Figure 8. Scatterplots of observed against predicted Category A response probabilities from RULEX and the GCM for Observers 1 and 2 of Nosofsky's (1986) experiment. Circles, dimensional; diamonds, interior–exterior; squares, criss-cross; triangles, diagonal.

mary fits for the GCM are also presented. Scatterplots of observed against predicted Category A response probabilities for each model are shown in Figure 8. The GCM yields a better fit than does RULEX for Observer 2's diagonal-categorization data. The models perform about equally for both observers in the dimensional categorization. In the remaining five cases, however, RULEX yields a better fit than does the GCM. The improvement in fit is substantial for the criss-cross and interior–exterior categorizations. It is also worth noting that the mean estimate of the perceptual-criterial noise parameter across the eight data sets was $\sigma = .63$. The large value of σ makes sense given that highly perceptually confusable stimuli were used.[7] Finally, note that we cannot evaluate RULEX's ability to predict the distribution of generalizations in this experiment because only 2 observers were tested.

Although RULEX compares favorably with the GCM on its ability to fit the classification transfer data, some caveats are in order. First, Nosofsky (1986) has already noted limitations on the ability of the GCM to handle these data, and there appear to be good psychological reasons why such limitations may have arisen. In Nosofsky's (1986) paradigm, the 2 observers classified the same transfer stimuli repeatedly during the test phase. (Multiple observations of each transfer stimulus were needed in order to obtain sufficiently large sample sizes for model fitting.) Unfortunately, it seems likely that once an observer makes an initial decision regarding the category membership of a transfer stimulus, this initial decision is likely to influence subsequent ones. For example, Nosofsky (1986) suggested that observers might augment their category representations with inferred exemplars. Once a transfer stimulus is classified into a category a given number of times, the observer might augment his or her category representation with this transfer stimulus. Thus, comparing RULEX to the standard GCM without making allowance for a memory-augmentation process is

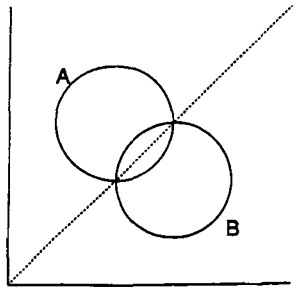

Figure 9. Contours of equal likelihood for the bivariate normal category distributions tested by Ashby and Maddox (1990). The diagonal line is the optimal decision boundary for partitioning the space into category regions.

probably not completely fair to the exemplar-based approach.

Second, Maddox and Ashby (1993, p. 62) have reported even better AIC fits of their quadratic decision boundary model (QDBM) to Nosofsky's (1986) data than those reported here for RULEX. According to the QDBM, the observer uses a decision boundary with a quadratic form to partition the space into response regions. The previously obtained fit values for the QDBM are reported along with those of the GCM and RULEX in Table 12. The QDBM gives a slightly worse AIC fit than does RULEX for Observer 1's dimensional and interior–exterior conditions data, is essentially the same as RULEX for Observer 2's dimensional condition data, but gives a better fit in the remaining five cases. A straightforward interpretation is that the ideas in RULEX are incorrect, and that observers instead use quadratic decision boundaries for dividing the perceptual space into response regions. It should be noted, however, that in each condition the quadratic model uses seven free parameters for fitting these data, so it has considerably more flexibility than does RULEX. Maddox and Ashby (1993, Figures 4 and 5) provided illustrations of the best-fitting quadratic boundaries for these studies. Inspection of these boundaries suggests that most of them have an overall form similar to the logical rule-boundaries posited in RULEX, but with some curvilinear distortion. One possibility is that the observers were trying to implement the logical rules assumed in RULEX, but had difficulty in maintaining a fixed criterion setting for their rule boundaries across the range of the perceptual space. For example, the precise criterion that an observer sets for using a rule based on angle might be influenced by the perceived size of the circle as well. Alternatively, perhaps the perceptual representations for these stimuli are more complex than the ones we assume for simplicity in this modeling, and the extra free parameters of the QDBM are capturing this complexity (cf. Maddox & Ashby, 1998). We conclude that our fits of RULEX to Nosofsky's (1986) categorization data show some real promise for the model, but it still has a way to go before it can match the precision achieved by the QDBM.

APPLYING RULEX TO LARGE-SIZE, PROBABILISTIC CATEGORY STRUCTURES

Thus far in our article, we have focused solely on experimental paradigms involving structures with a small number of training exemplars that are assigned deterministically by the experimenter to the alternative categories. A much different type of experimental paradigm involves structures with a large number of training exemplars that are assigned probabilistically to the alternative categories. In particular, Ashby, Maddox, and their colleagues have often used an experimental paradigm in which the categories are defined by bivariate normal distributions (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Maddox & Ashby, 1993; see also McKinley & Nosofsky, 1995, 1996). An interesting challenge for RULEX is whether or not it can account for performance in such tasks. In this final section of this article, our purpose is to provide some preliminary ideas along these lines.

Consider the category structure illustrated in Figure 9. There are two categories defined by bivariate normal distributions. Each distribution is represented schematically by a contour of equal likelihood (Ashby & Gott, 1988), which is a locus of points that are equally likely to be produced by the distribution. For normal distributions, these contours are always circular or elliptical in shape. The center of each ellipse gives the mean of the normal distribution on each of its dimensions. The expanse of the ellipse along each dimension represents the variability of the distribution along that dimension. The correlation between the dimensions is represented by the angle of orientation of the ellipse. Figure 9 illustrates a simple situation in which both category distributions have equal variance along both dimensions and where there is zero correlation.

In the usual version of the paradigm, each normal distribution defines a category. Thus, on each trial of the experiment: (1) a category distribution is selected; (2) an exemplar from this distribution is randomly chosen and presented to the observer; (3) the observer guesses the exemplar's category assignment; and (4) corrective feedback is then provided. By the time learning is completed, an observer may have experienced thousands of unique training exemplars from each category. Because the distributions are overlapping, it is impossible to classify all exemplars perfectly. However, it is possible to define an optimal decision boundary that maximizes classification accuracy. It is well known that for bivariate normal distributions, the optimal boundary is always linear or quadratic in form. Indeed, for the situation illustrated in Figure 9, the optimal decision boundary is simply the 45° line running through the space. Any point falling to the upper left of the line should be classified in Category A, and any point falling to the lower right should be classified in Category B.

Ashby, Maddox, and their colleagues have provided evidence that in a variety of situations like those illustrated in Figure 9, observers do indeed appear to adopt a boundary

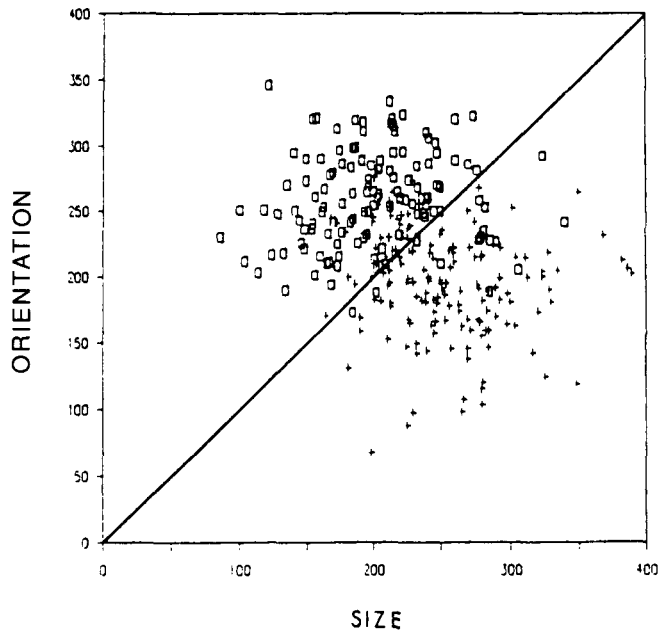

Figure 10. Category responses made by a representative observer from Ashby and Maddox's (1990) experiment using the category structure illustrated in Figure 9. Squares, Category A responses; crosses, Category B responses. From "Integrating Information From Separable Psychological Dimensions," by F. G. Ashby and W. T. Maddox 1990, *Journal of Experimental Psychology: Human Perception & Performance, 16*, p. 608 (Figure 6). Copyright 1990 by the American Psychological Association. Reprinted with permission.

with the same form as the optimal boundary (following sufficient experience with the training exemplars). An example from Ashby and Maddox (1990) is shown in Figure 10. In the figure, locations marked with a square represent items for which the observer made a Category A response and locations marked with a cross represent Category B responses. The figure illustrates that the optimal boundary neatly separates the space into the two response regions, lending support to the idea that the observer used this diagonal linear boundary for making his or her responses. (Occasional inconsistent responses across the boundary are attributed to the effects of perceptual and criterial noise.)

These results appear to pose a severe challenge to RULEX. The key idea in RULEX is that observers establish decision boundaries that are orthogonal to the coordinate axes and not at oblique angles. Clearly, there is no small set of orthogonal decision boundaries that could produce the response pattern in Figure 10. The question arises, however, of what RULEX would predict if an orthogonal single-dimension boundary were supplemented with exceptions.

Suppose that an observer established an orthogonal boundary based on Dimension 1 values, such as illustrated by the dashed line in Figure 11. The classification responses predicted by the orthogonal boundary agree with those predicted by the optimal diagonal linear boundary (the dotted line), except for the regions marked Ax and Bx in the figure. During training, observers would experience a great many training exemplars in these regions that are misclassified by the orthogonal boundary. Following in the spirit of the RULEX model, it is reasonable to posit that an observer might store a representative exception (or small number of exceptions) in each of these regions and use them to supplement the single-dimension rule (i.e., store an A exception in region Ax and a B exception in region Bx). Applying the process formalized in Equations 1–5, the idea would be that an observer would classify an item based on the exceptions if the item were sufficiently similar to the exceptions; otherwise, the rule would be used.

Figure 12 shows the classification predictions made by a single simulation of RULEX in a situation in which we treated the locations of the exceptions as free parameters (a single exception was assumed for each region). The locations of the exceptions were chosen to maximize the agreement of RULEX with the predictions made by the optimal linear boundary.[8] It is evident from inspection that the responses predicted by this version of RULEX are in close accord with the responses defined by a diagonal linear decision boundary, with all stimuli falling to the upper left of the boundary classified in Category A and all stimuli falling to the lower right classified in Category B. One way of thinking about this demonstration is that it illustrates a situation in which an optimal diagonal decision boundary can be approximated by a simple one-dimensional rule together with a couple of exceptions.

It is critical to understand, however, that the overlap between the optimal diagonal linear boundary and the

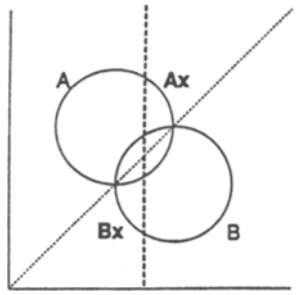

Figure 11. Illustration of a rule-plus-exception strategy for classifying objects from the Figure 9 category structure. The dashed line indicates the dimensional rule, Ax and Bx indicate the exception regions. The dotted diagonal boundary is shown for purposes of comparison.

RULEX boundary arises only in the region of the stimulus space in which observers have experienced training exemplars. If one tested observers with unfamiliar transfer items located far from the original training region, the patterns of generalization predicted by the two models would differ dramatically. In particular, according to RULEX, if an unfamiliar transfer item is presented that is dissimilar to the stored training exceptions, the exception-use process will no longer operate, so classification decisions will again be based on dimensional rules that are orthogonal to the coordinate axes. The result is that although the RULEX decision boundary is essentially linear in the local training region, it ends up being globally nonlinear when viewed across the entire span of the stimulus space. Thus, this question of how observers generalize in unfamiliar regions is a critically important one that needs to be investigated in future work.

Another important question that arises is why observers would first form an orthogonal linear boundary and then later "patch it up" with exceptions. For example, in the Figure 9 structure, observers could simply store the central tendency or "prototype" of each category and classify on the basis of similarity to the prototype. Objects more similar to the A prototype would be classified in Category A, and likewise for Category B. Such a classification strategy would also produce the diagonal linear boundary illustrated in the figure. However, forming each category prototype requires averaging over a highly diffuse set of exemplars spanning the range of the psychological space. By contrast, forming exceptions in the local regions marked Ax and Bx may be less cognitively demanding: It seems easier to form a summary representation for highly similar exemplars in a local region of space than for highly disparate exemplars spanning a wide region.

If these ideas involving RULEX are correct, then one straightforward prediction is that at very early stages of learning with normally distributed categories, the decision boundaries that best describe observers' behavior should be orthogonal to the coordinate axes rather than at

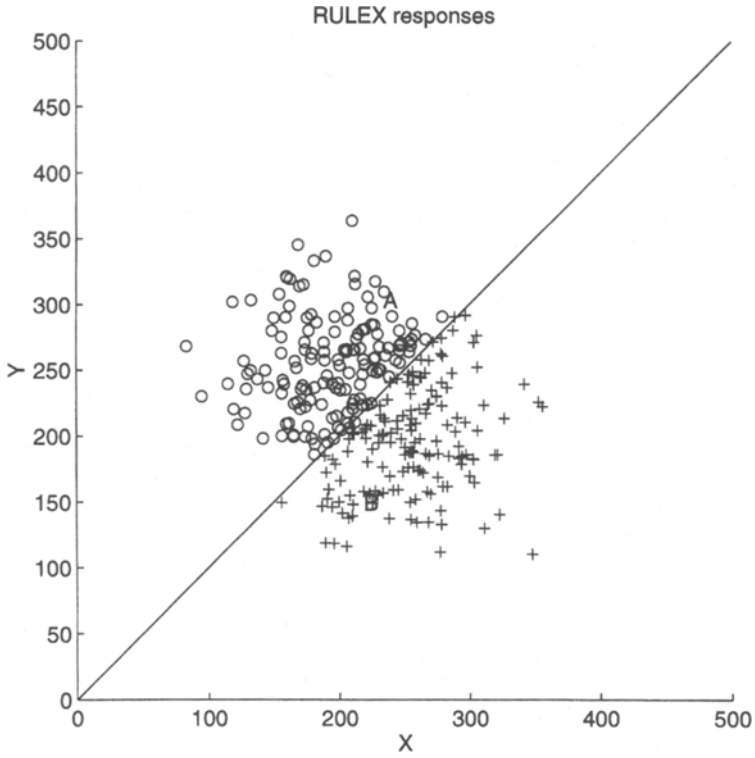

Figure 12. Classification responses predicted from a simulation of the RULEX strategy illustrated in Figure 11. Circles, Category A responses; crosses, Category B responses. The locations of the A and B exceptions are also shown in the figure. The solid diagonal boundary is shown for purposes of comparison.

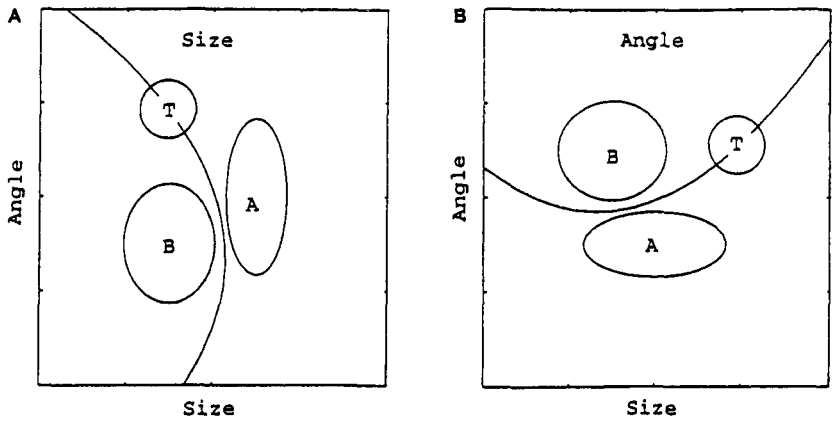

Figure 13. Category structures tested by McKinley and Nosofsky (1996, Experiment 2). Each bivariate normal category distribution is represented by a contour of equal likelihood, as explained in the text. From "Selective Attention and the Formation of Linear Decision Boundaries," by S. C. McKinley and R. M. Nosofsky, 1996, *Journal of Experimental Psychology: Human Perception & Performance, 22*, p. 307 (Figure 6). Copyright 1996 by the American Psychological Association. Adapted with permission.

oblique angles. Alfonso-Reese (1996) has reported some data that are consistent with this prediction. She tested 10 individual observers in a probabilistic classification paradigm using two bivariate normal category distributions, where the stimuli were lines varying in their lengths and angles of orientation. At various points during the learning phase, she tested observers in a transfer phase in which objects spanning the two-dimensional space were presented for classification without feedback. The goal was to investigate the types of decision boundaries that observers had adopted at various points during learning. (If one continued to provide feedback, then the decision boundary that an observer had adopted would undergo continual change, so a single "snapshot" could never be taken.) Alfonso-Reese found that, during the early stages of learning (e.g., following 10 and 30 training trials), the linear decision boundary that provided the best account of her observers' response patterns was usually a nearly orthogonal linear boundary. For most of the observers she tested, the linear boundary that provided the best fit to the early transfer data was closer in slope to an orthogonal linear boundary than to the diagonal linear boundary that would optimize performance. These results support the RULEX prediction that, even in probabilistic classification paradigms involving large numbers of training exemplars, observers initiate their learning by searching for single-dimension rules that best partition the objects into categories.

The important role of orthogonal linear boundaries in probabilistic classification designs has also been brought out by the work of McKinley and Nosofsky (1996, Experiment 2). Twenty-four individual observers were tested in the two categorization conditions illustrated in Figures 13A and 13B. The stimuli were circles varying in their sizes and angles of orientation of a radial line. In a training phase, the observers learned to classify the stimuli into two bivariate normal categories, labeled A and B in Figures 13A and 13B. The structure of these normal distributions is illustrated in the figures in terms of their contours of equal likelihood. Following the training phase, a test phase was conducted in which observers classified stimuli from the original training distributions as well as a new transfer region, marked "T" in the figures. The transfer region was used to provide diagnostic information concerning the types of decision boundaries that observers were using to partition the space into categories.

As can be seen in Figures 13A and 13B, in the size categorization, the dimension of size was primarily relevant for performing the classification, whereas in the angle categorization, the dimension of angle was primarily relevant. Indeed, observers could achieve nearly optimal performance in this design by using a single-dimension rule in each condition—that is, by using a linear boundary orthogonal to the size dimension in the size categorization, and likewise for the angle categorization. (Systematic exceptions to these rules are rarely experienced, so observers might never store any exceptions.) However, in both conditions, a quadratic decision boundary provides the ideal-observer boundary for partitioning the space into categories. These ideal-observer boundaries are illustrated along with the category distributions in Figures 13A and 13B. Note that the ideal-observer boundaries and the single-dimension rule boundaries make dramatically different predictions regarding how observers will classify objects in the transfer regions. The single-dimension rule boundary predicts that the transfer items will be classified in Category B with high probability, whereas the ideal-observer boundary predicts that the transfer items will be classified into Categories A and B with roughly equal probability.

Despite the fact that a quadratic decision boundary is the ideal-observer boundary, McKinley and Nosofsky (1996) found that, in the size and angle conditions, linear decision boundaries actually provided slightly better

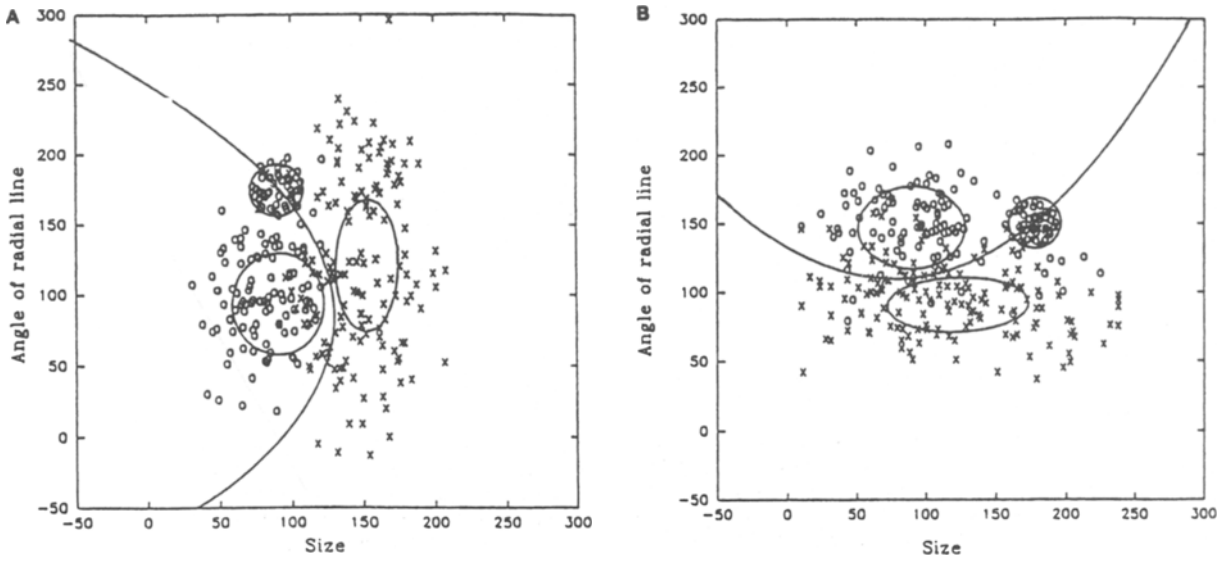

Figure 14. Response patterns for two representative subjects from the size and angle conditions illustrated in Figure 13. Category A responses are labeled ×, and Category B responses are labeled ○. From "Selective Attention and the Formation of Linear Decision Boundaries," by S. C. McKinley and R. M. Nosfosky, 1996, *Journal of Experimental Psychology: Human Perception & Performance, 22*, p. 309 (Figure 7). Copyright 1996 by the American Psychological Association. Adapted with permission.

AIC fits to the subjects' classification response patterns than did general quadratic boundaries, and far outperformed the ideal-observer boundary. Furthermore, in most cases, the best-fitting linear boundary was essentially orthogonal to the relevant dimension in these conditions. The response patterns for two representative observers from the size and angle conditions are illustrated in Figure 14, where Category A responses are labeled × and Category B responses are labeled ○. Inspection of the figure indicates clearly that both observers overwhelmingly classified objects in the transfer regions into Category B, which is consistent with the idea that they formed a nearly orthogonal linear boundary along the relevant dimension for partitioning the response regions of Categories A and B. Averaged across all observers, the probability with which patterns in the transfer regions were classified into Category B was .92 in the size condition and .84 in the angle condition, consistent with the hypothesis that many of the individual observers used the single-dimension rules.

In summary, McKinley and Nosofsky's (1996) results support the RULEX prediction that people may have strong tendencies to form orthogonal linear boundaries for purposes of classification, even in designs involving normally distributed category structures in which the form of the ideal-observer boundary is highly nonlinear.

GENERAL DISCUSSION

In summary, in this article we have formalized a rule-plus-exception (RULEX) model of how observers classify objects in continuous, multidimensional spaces. At the heart of the model is the assumption that observers partition continuous multidimensional spaces into category regions by forming decision boundaries that are orthogonal to the coordinate axes. If needed, the observers then remember occasional exceptions to these single-dimension rules. Objects that are sufficiently similar to an exception are classified in the category to which the exception belongs, otherwise the rules are applied.

In the initial tests presented in this article, we demonstrated that the continuous-dimension RULEX model was capable of providing excellent quantitative fits to previously reported sets of classification data as well as to some new data sets reported herein. In addition to predicting extremely accurately the averaged classification probabilities for individual stimuli at time of transfer, a major accomplishment of RULEX is that it also describes well the heterogeneity in classification response patterns seen at the individual-observer level. Specifically, RULEX provides simultaneous good fits to averaged classification data and to the distributions of generalizations displayed by individual observers that underlie these averaged data. More complex logical rules, such as conjunctive, biconditional, and extreme-value rules, can be formed by combining orthogonal boundaries along multiple dimensions, and we demonstrated reasonably good fits of such models to individual-observer classification data reported by Nosofsky (1986). Finally, we provided some ideas about how RULEX may be applicable in probabilistic classification designs involving normally distributed categories and reviewed some preliminary evidence in favor of these ideas.

One of the major directions for future research is to develop a learning version of the continuous-dimension RULEX model. In its current form, the model is intended to describe only asymptotic performance observed at the completion of training. It is critical to understand, how-

ever, the processes that lead observers to adopt certain classification rules rather than others, as well as how exceptions come to be formed. In addition to describing the category-learning process, a learning version of RULEX may allow for a reduction in the number of free parameters currently required for fitting the model to data. Instead of estimating the rule probabilities (Equation 5) on the basis of best fits to the transfer data, the learning model might allow for the a priori prediction of these rule probabilities. Our demonstrations that the RULEX model is capable of yielding extremely accurate quantitative predictions of individual observers' classification response patterns strongly encourages future attempts to develop such a learning model.

Although our emphasis in this article has been on the role that RULEX processes play in classification, we do not take the position that stored exemplars play no role in people's category representations. Indeed, there is good evidence from previous research that experiences with specific exemplars exert an important influence on classification behavior, even in situations in which simple rules are available for performing a task. For example, Nosofsky (1991) conducted experiments in which subjects learned classifications defined by exceedingly simple logical rules. The frequency with which observers experienced different exemplars that satisfied those rules was manipulated. Nosofsky (1991) found that observers' goodness-of-example judgments and speeded classifications were strongly influenced by these frequency manipulations in a manner consistent with the idea that stored exemplars formed part of the category representation. Also, in Nosofsky, Palmeri, and McKinley's (1994) tests of the binary-valued version of RULEX, it was discovered that a complete account of the distribution-of-generalizations data required recourse to the idea that at least some exemplar-based classification had taken place. Even in experiments in which observers are given explicit instructions to use certain logical rules for purposes of classification, there is evidence that the specific exemplars on which the observers are trained influence both their categorization and old–new recognition judgments (e.g., Brooks, Norman, & Allen, 1991; Nosofsky et al., 1989; Palmeri & Nosofsky, 1995).

In addition, the ability of an observer to implement an orthogonal linear boundary, such as assumed in RULEX, is strongly influenced by the types of dimensions that compose the objects. Forming an orthogonal linear boundary along a single dimension would seem to require the ability to "selectively attend" to that dimension while ignoring values along other dimensions. Therefore, we focused throughout our article on stimuli varying along highly separable dimensions. Such dimensions are ones that remain psychologically distinct when in combination, and where selective attention can operate with high efficiency (e.g., Garner, 1974; Shepard, 1964). By contrast, integral dimensions are ones that combine into relatively unanalyzable wholes, and where selective attention is difficult. Good examples are colors varying in brightness and saturation, or tones varying in loudness and pitch. Nosofsky (1987, 1998) and McKinley and Nosofsky (1996) provided evidence that when observers learn to classify integral-dimension stimuli, they fail to form orthogonal linear boundaries along single dimensions, even when such boundaries would produce nearly optimal performance. Instead, the patterns of performance observed under such conditions are more consistent with the idea that similarity comparisons to stored exemplars drive classification (McKinley & Nosofsky, 1996; Nosofsky & Palmeri, 1997).

Therefore, we believe that a full account of perceptual categorization will involve the development of a hybrid model that makes use of both RULEX strategies together with exemplar storage (see also J. R. Anderson, Kline, & Beasley, 1979; Nosofsky et al., 1989; Nosofsky & Palmeri, 1997; Nosofsky, Palmeri, & McKinley, 1994; Palmeri, 1997). Indeed, some important developments along these lines are currently taking place. For example, Vandierendonck (1995) proposed a parallel rule activation and rule synthesis (PRAS) model in which both rules and exemplars are represented in a common production-system framework. Rather than abstracting rules along just single dimensions, however, in PRAS the system abstracts rules corresponding to rectangular regions in psychological space that are defined by pairs of exemplars. To date, however, Vandierendonck has applied the model only in a limited situation in which observers learned two categories defined by just two exemplars each, so the generality with which such a rule-abstraction mechanism may operate remains unknown.

Erickson and Kruschke (1998) proposed a connectionist model called ATRIUM (attention to rules and instances in a unified model) that consists of two modules. One module classifies objects through the use of single-dimension rules, as we have assumed in RULEX. A second module learns associations between exemplars and categories. (The second module is Kruschke's, 1992, ALCOVE model, which incorporates key components of Nosofsky's, 1986, exemplar-based GCM within a connectionist framework.) The hybrid system learns which module is superior for classifying exemplars in different regions of the psychological space. In essence, it learns to pay special attention to exemplars that are exceptions to the single-dimension rule and to use the exemplar-based module for classifying these objects.

We do not view Erickson and Kruschke's (1998) modeling with ATRIUM as being in competition with our own, but rather as pursuing complementary research strategies. Erickson and Kruschke have developed a specific learning model involving principles of both rule formation and exemplar storage, and have tested it on its ability to predict averaged learning data in category structures in which all observers are expected to learn the same simple rules. By contrast, we have bypassed questions about learning in this research, and have investigated instead the extent to which RULEX processes may underlie performance in diverse tasks, including ones in which multiple RULEX strategies are available. We have further focused on the ability of a RULEX model to ac-

count for the dramatic individual differences in patterns of generalization that underlie the averaged classification data. Although the paths that we are taking are somewhat different, they should converge to similar models once we extend RULEX with learning principles combined with exemplar storage and Erickson and Kruschke (1998) extend ATRIUM with multiple-rule modules and stochastic rule-selection mechanisms.

Another model under current development that is related to RULEX is Ashby, Alfonso-Reese, and Turken's (1995) COVIS (competition between verbal and implicit systems) model. According to COVIS, classification performance is mediated by two separate systems. One system learns simple "verbal rules," similar to the types of rules formed by RULEX. A second system learns more complex decision boundaries for partitioning a space into categories, and these complex boundaries are difficult to verbalize. Classification performance is conceptualized as emerging from a competition between the verbal-rule system and the complex decision-boundary system. A point that distinguishes COVIS from RULEX and ATRIUM is that there is no form of exemplar storage and retrieval in COVIS. An interesting challenge for this model is to explain the effects of exemplar frequency on observers' classification judgments (e.g., Nosofsky, 1991), as well as why observers show enhanced recognition memory for exceptions to category rules (Palmeri & Nosofsky, 1995). Nevertheless, the aspect of COVIS that places emphasis on the development of simple verbal rules as a key component of categorization converges with our own ideas about the importance of single-dimension rules in RULEX.

Finally, the idea that human judgments may often rely primarily on information provided along single dimensions has also taken hold in a modern theory of inductive reasoning. Gigerenzer and Goldstein (1996) considered inferential tasks in which observers make a choice between two alternatives on a quantitative dimension. The task requires that the inference be based on information stored in memory. For example, an observer might be presented with the question: "Which city has a larger population? (a) Hamburg (b) Potsdam." Gigerenzer and Goldstein proposed a family of algorithms for making such choices based on "one-reason decision making." In these algorithms, a single cue providing probabilistic information concerning the answer is accessed from memory and is used to draw the inference. For example, an observer might remember that Hamburg has a professional soccer team, whereas Potsdam does not, and use this single-cue information to infer that Hamburg has a larger population than does Potsdam. Gigerenzer and Goldstein reviewed evidence suggesting that these one-reason decision-making algorithms provide accurate descriptions of human performance in these tasks. In addition, they documented that these one-reason algorithms yielded as many correct inferences about unknown features of real-world environments as did a variety of information-integration algorithms.

We find the parallels between RULEX and these one-reason decision-making algorithms to be striking. Both models suggest that human observers may place primary reliance on information from single dimensions for making categorization or inductive-reasoning judgments, and both models can often yield performance levels that approximate that of an optimal decision maker. We find it intriguing that the principles underlying single-dimension RULEX strategies in categorization may generalize to other cognitive domains.

REFERENCES

AHN, W.-K., & MEDIN, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, **16**, 81-122.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

ALFONSO-REESE, L. A. (1996). *Dynamics of category learning*. Unpublished doctoral dissertation, University of California, Santa Barbara.

ANDERSON, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, **98**, 409-429.

ANDERSON, J. R., KLINE, P. J., & BEASLEY, C. M. (1979). A general learning theory and its application to schema abstraction. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 13, pp. 277-318). San Diego, CA: Academic Press.

ASHBY, F. G., ALFONSO-REESE, L., & TURKEN, U. (1995, November). *Competition between verbal and implicit rules of category learning*. Paper presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles.

ASHBY, F. G., & GOTT, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.

ASHBY, F. G., & LEE, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, **120**, 150-172.

ASHBY, F. G., & MADDOX, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 598-612.

ASHBY, F. G., & MADDOX, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 50-71.

ASHBY, F. G., & MADDOX, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

ASHBY, F. G., MADDOX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144-151.

ASHBY, F. G., & TOWNSEND, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154-179.

BARSALOU, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 629-654.

BOURNE, L. E., JR. (1970). Knowing and using concepts. *Psychological Review*, **77**, 546-556.

BROOKS, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

BROOKS, L. R., NORMAN, G. R., & ALLEN, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, **120**, 278-287.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. (1956). *A study of thinking*. New York: Wiley.

ENNIS, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, **117**, 408-411.

ERICKSON, M. A., & KRUSCHKE, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, **127**, 107-140.

ESTES, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

ESTES, W. K., & MADDOX, W. T. (1995). Interactions of similarity, base rate, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1075-1095.

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

GIGERENZER, G., & GOLDSTEIN, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, **103**, 650-669.

HAYES-ROTH, B., & HAYES-ROTH, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning & Verbal Behavior*, **16**, 321-328.

HEIT, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1264-1282.

HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

HUNT, E. B., MARIN, J., & STONE, P. J. (1966). *Experiments in induction*. New York: Academic Press.

KNAPP, A. G., & ANDERSON, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 616-637.

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

KRUSKAL, J. B., & WISH, M. (1978). *Multidimensional scaling*. London: Sage Publications.

LAMBERTS, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1003-1021.

LEVINE, M. (1975). *A cognitive theory of learning: Research on hypothesis testing*. Hillsdale, NJ: Erlbaum.

MADDOX, W. T., & ASHBY, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, **53**, 49-70.

MADDOX, W. T., & ASHBY, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 301-321.

MARTIN, R. C., & CARAMAZZA, A. (1980). Classification in well-defined and ill-defined categories: Evidence for common processing strategies. *Journal of Experimental Psychology: General*, **109**, 320-353.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 128-148.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 294-317.

MEDIN, D. L., ALTOM, M. W., EDELSON, S. M., & FREKO, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 37-50.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

MEDIN, D. L., & SCHWANENFLUGEL, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 355-368.

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

MEDIN, D. L., WATTENMAKER, W. D., & MICHALSKI, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, **11**, 299-339.

MYUNG, I. J., & PITT, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

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

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

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

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

NOSOFSKY, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.

NOSOFSKY, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, **19**, 131-150.

NOSOFSKY, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, **43**, 25-53.

NOSOFSKY, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 322-339.

NOSOFSKY, R. M., CLARK, S. E., & SHIN, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 282-304.

NOSOFSKY, R. M., GLUCK, M. A., PALMERI, T. J., MCKINLEY, S. C., & GLAUTHIER, P. T. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, **22**, 352-369.

NOSOFSKY, R. M., & PALMERI, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, **104**, 266-300.

NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.

PALMERI, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 324-354.

PALMERI, T. J., & NOSOFSKY, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 548-568.

PAVEL, M., GLUCK, M. A., & HENKLE, V. (1988). Generalization by humans and multi-layer networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 680-687). Hillsdale, NJ: Erlbaum.

POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

REED, S. K. (1996). *Cognition: Theory and applications*. Pacific Grove, CA: Brooks/Cole.

RESTLE, F. (1962). The selection of strategies in cue learning. *Psychological Review*, **69**, 329-343.

RIPS, L. J., SCHOBEN, E. J., & SMITH, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning & Verbal Behavior*, **12**, 1-20.

ROSCH, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

ROSCH, E., & MERVIS, C. B. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

ROSCH, E., SIMPSON, C., & MILLER, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 491-502.

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

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

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

SHEPARD, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. Pomerantz & G. Lockhead (Eds.), *Perception of structure* (pp. 53-71). Washington, DC: American Psychological Association.

SHEPARD, R. N., HOVLAND, C. I., & JENKINS, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75** (13, Whole No. 517).

SMITH, E. E., & MEDIN, D. L. (1982). *Categories and concepts*. Cambridge, MA: Harvard University Press.

SMITH, J. D., MURRAY, M. J., & MINDA, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 659-680.

TRABASSO, T., & BOWER, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.

VANDIERENDONCK, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, **2**, 442-459.

WARD, T. B., & SCOTT, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, **15**, 42-54.

NOTES

1. For simplicity, we assume that all stimuli give rise to normally distributed perceptual representations, that they have the same perceptual variance, σ_p^2, on each of their dimensions, and that there is zero covariance. In addition, each decision boundary has constant criterial variance, σ_c^2. Under these assumptions, the value of σ used for computing the relevant classification probabilities is $\sigma = \sqrt{(\sigma_p^2 + \sigma_c^2)}$. See Ashby and Maddox (1993) for further details regarding the parameters in the decision-boundary models.

2. For the present categorization transfer data, the value of ln L is given by

$$\ln L = \Sigma M_i! - \Sigma\Sigma f_{ij}! + \Sigma\Sigma f_{ij} \cdot \ln(p_{ij}),$$

where M_i is the observed frequency of stimulus i, f_{ij} is the observed frequency with which stimulus i was classified in category j, and p_{ij} is the predicted probability from the model that stimulus i is classified in category j. This likelihood function assumes that the category responses for each stimulus are multinomially distributed and that the response distributions are independent.

3. Our fit comparisons involving AIC are used as a preliminary guide but should be interpreted with caution. First, it seems likely that with sufficiently large sample sizes, the AIC measure will tend to favor a higher parameter model over a lower parameter one, because the extra free parameters are useful in accounting for the sundry "noise" factors that are not part of the models' approximations. Second, even with the number of free parameters held fixed, certain models have functional forms that are inherently more flexible than those of alternative models, but the AIC statistic is insensitive to this factor (Myung & Pitt, 1997). Third, the AIC statistic does not take into account the flexibility arising from model-selection processes, in which a family of models actually exists but one particular candidate from the family is eventually chosen as the representative. (We should note that, in our case, the decision to restrict RULEX to the set of all rules that had no more than two exceptions was made a priori without consideration of the data we were fitting. In our design, rules with three or more exceptions produce, at most, a 57% accuracy rate, which is nearly chance.) Alternative measures of model fit that are sensitive to these factors are currently being investigated (e.g., Myung & Pitt, 1997), but are at too early a stage of development to be used in the present research.

4. In the context of criticizing an article by McKinley and Nosofsky (1996), Maddox and Ashby (1998) have also argued that model fits involving only averaged data can be misleading. While expressing fundamental agreement with this general point, Nosofsky (1998) replied that the particular criticisms raised by Maddox and Ashby (1998) were misguided. See Maddox and Ashby (1998) and Nosofsky (1998) for details regarding this particular debate.

5. Note that each observer classified each transfer stimulus in three separate blocks. In computing the observed distribution of generalizations, an observer is defined as classifying an object in Category A if he or she classified it in Category A in at least two of the three blocks. See Appendix A for an explanation of how RULEX is used to predict the distribution of generalizations.

6. The method for fitting RULEX to the diagonal categorization is the same as that described previously in the text, except that there are modifications in the exact form of the similarity and distance functions that govern the exception-use process (Equation 2). In particular, distance in the psychological space is computed by using a Euclidean metric instead of a city-block metric, and similarity is related to distance by a Gaussian function instead of by an exponential decay function. Nosofsky (1985, 1986) found that these functions provided better accounts of overall similarity relations among these highly confusable stimuli than did the standard functions. Ennis (1988) subsequently demonstrated that these alternative functions might simply be reflecting extensive Gaussian noise in the internal perceptual representations of the stimuli. Rather than modeling the Gaussian noise explicitly, however, we use the alternative distance metric and similarity function for simplicity in the model fitting.

7. The best-fitting parameters for each individual condition are available from the first author upon request.

8. Because the linear boundary is the optimal classification strategy for this category structure, maximizing the agreement of RULEX with the linear boundary is basically the same as finding the parameters that would allow RULEX to maximize performance. It seems reasonable that adaptive learning mechanisms may exist that would place exception representations at locations that would maximize performance. In applying RULEX to predict the results in Figure 12, we used a deterministic response rule for applying exceptions (i.e., we used deterministic versions of Equations 1 and 3), and also assumed zero perceptual-criterial noise in applying the single-dimension rule. If these assumptions were relaxed, then there would be occasional scatter of responses across the 45° linear boundary, as observed by Ashby and Maddox (1990).

APPENDIX A
Using RULEX to Predict the Distribution of Generalizations

RULEX is used to predict the distribution of generalizations as follows. First, for any given RULEX strategy R_j, one uses Equations 1–5 to predict the probability that stimulus i is classified in Category A on any given trial, $P(A \mid i, R_j)$. By definition, stimulus i is classified in Category A during the transfer phase if an observer classifies it in Category A on at least a majority, M, of the test blocks. Thus, assuming N test blocks, it follows by using the binomial expansion that the predicted probability that stimulus i is classified in Category A during transfer, given RULEX strategy R_j, $P_T(A \mid i, R_j)$, is given by

$$P_T(A \mid i, R_j) = \sum_{m=M}^{N} \binom{N}{m} P(A \mid i, R_j)^m \cdot P(B \mid i, R_j)^{N-m}. \quad (A1)$$

The probability of observing generalization profile G, given RULEX strategy R_j, is then given by

$$P(G \mid R_j) = \prod P_T(A \mid i, R_j)^{\delta(i)} \cdot P_T(B \mid i, R_j)^{1-\delta(i)}, \quad (A2)$$

where $\delta(i) = 1$ if profile G has response A in position i, and $\delta(i) = 0$ if profile G has response B in position i. Finally, the overall probability of profile G is found by summing these conditional probabilities weighted by the estimated probability of RULEX strategy R_j:

$$P(G) = \sum P(G \mid R_j) \cdot P(R_j). \quad (A3)$$

APPENDIX B
Classification Transfer Data and Distribution-of-Generalizations Data for the Nonlearners from Experiment 2

Table B1 reports the classification transfer data obtained for 44 of the 65 nonlearners from Experiment 2. The 44 nonlearners that are included made 10 or fewer errors during the final 35 training trials of the experiment. (We do not report the data of the 21 additional nonlearners who made greater than 10 errors, as most of these individuals were performing at levels close to chance.) The distribution-of-generalizations data from the 44 nonlearners are reported in Table B2. Only the generalization profiles displayed by at least 2 observers are reported individually. Besides displaying lower average accuracy on the training exemplars than did the learners, the nonlearners' performance differed from that of the learners in two main respects. First, the learners had a strong tendency to classify Transfer Stimuli 5 and 9 into Category A, whereas the nonlearners had a slight tendency to classify these transfer stimuli into Category B. Second, the learners had classified Transfer Stimuli 15 and 16 into Categories A and B with roughly equal probability, whereas the nonlearners tended to classify these stimuli into Category A. We fitted RULEX to the composite data of the nonlearners using the methods described in the main text. These fits need to be interpreted with a good deal of caution, because, as currently formalized, the model applies only to observers who have formed rules and exceptions that accurately solve the problem. Nevertheless, the fit results suggested that the nonlearners used broader generalization gradients when applying the exceptions than did the learners. Also, like the learners, the nonlearners made predominant use of Dimension 2 (brightness) for forming their rules, but this pattern was not as extreme for the nonlearners as for the learners.

Table B1
Observed Classification Transfer Data of the Nonlearners From the Final Set of Test Blocks in Experiment 2

Stimulus	Category A Response Probability
B1	.061
2	.220
B3	.152
4	.091
5	.402
A6	.644
7	.485
B8	.227
9	.417
A10	.773
A11	.864
12	.742
B13	.227
14	.553
15	.780
16	.652

Note—Training stimuli from Categories A and B are denoted with an A or a B, respectively.

Table B2
Observed Distribution of Generalization Data of the Nonlearners From the Final Set of Test Blocks in Experiment 2

Profile	Probability
BBBBBAAAA	.045
BBBBAAAAA	.068
BBBAAABBB	.068
BBABBAAAA	.091
BBAABAAAA	.045
Other	.683

(Manuscript received March 13, 1997;
revision accepted for publication January 7, 1998.)