Psychological Review
1994, Vol. 101, No. 1, 53-79

0033-295X/94/$3.00

Rule-Plus-Exception Model of Classification Learning

Robert M. Nosofsky, Thomas J. Palmeri, and Stephen C. McKinley

The authors propose a rule-plus-exception model (RULEX) of classification learning. According to RULEX, people learn to classify objects by forming simple logical rules and remembering occasional exceptions to those rules. Because the learning process in RULEX is stochastic, the model predicts that individual Ss will vary greatly in the particular rules that are formed and the exceptions that are stored. Averaged classification data are presumed to represent mixtures of these highly idiosyncratic rules and exceptions. RULEX accounts for numerous fundamental classification phenomena, including prototype and specific exemplar effects, sensitivity to correlational information, difficulty of learning linearly separable versus nonlinearly separable categories, selective attention effects, and difficulty of learning concepts with rules of differing complexity. RULEX also predicts distributions of generalization patterns observed at the individual subject level.

Psychologists have witnessed a major shift in the study of category learning during the past few decades. Early research was dominated by the concept-identification paradigm, in which subjects learned well-defined categories structured according to simple logical rules. Owing to the influence of researchers such as Rosch (1973) and Posner and Keele (1968), interest shifted to more ill-defined categories as might be found in the natural world. For ill-defined categories, no simple logical rules exist for classifying objects, and the boundaries demarcating alternative categories are fuzzy.

With the shift in emphasis from well-defined to ill-defined categories, there has also been a major shift in the types of models used for explaining classification learning. Early research was dominated by hypothesis-testing and rule-formation models (e.g., Bruner, Goodnow, & Austin, 1956; Hunt, Marin, & Stone, 1966; Levine, 1975; Neisser & Weene, 1962; Restle, 1962; Trabasso & Bower, 1968). Subjects were presumed to formulate and test simple hypotheses concerning the logical rules that defined category membership. By the time learning was completed, the category representation was presumed to consist of whatever simple logical rule partitioned the objects. In contrast, because no simple rules exist for classifying members of ill-defined category structures, hypothesis-testing models have largely fallen from the scene in psychological research and have been replaced by alternatives such as exemplar, Bayesian, and connectionist/distributed memory models.

These modern category learning models differ dramatically from classic hypothesis-testing models in their information-processing requirements. A common thread connecting most of the successful and well-known current models is that a great deal of information from the originally presented exemplars is retained in the category representation. According to exemplar models (e.g., Estes, 1986a; Hintzman, 1986; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986), category representations consist of the storage of all previously presented exemplars, and classification decisions involve massive similarity computations performed over these stored exemplars. A reasonable process interpretation for Anderson's (1990, 1991) rational model is that highly similar exemplars are grouped into clusters during the learning process, but this model retains the assumption that accurate statistical records are maintained over all the dimensions that compose the exemplars. Furthermore, in application to many classification learning paradigms, the best fitting version of the rational model creates a separate cluster for virtually every exemplar, so the model reduces essentially to a pure exemplar model (Anderson, 1990; Nosofsky, 1991a). In addition, according to the property set model of Hayes-Roth and Hayes-Roth (1977) and the configural cue model of Gluck and Bower (1988), there are records not only of all exemplars presented during learning but also of all lower order configurations of features that compose these exemplars. Such models presume that there is a massive amount of information about the presented exemplars that is retained in memory and used for making classification decisions.

Although the models just discussed have provided impressive accounts of a wide array of categorization phenomena, it is reasonable to question the plausibility of exemplar storage processes and the vast memory resources that they seem to require. Indeed, despite the shift away from formal hypothesis-testing models, the view that people might represent categories in terms of fairly simple logical rules remains stubborn (e.g., Martin & Caramazza, 1980; Medin, 1986; Nosofsky, Clark, & Shin, 1989; Pavel, Gluck, & Henkle, 1988; Ward & Scott, 1987). A key idea, according to such a view, is that, although ill-defined category structures cannot be learned by forming simple logical rules, perhaps such structures are learned by forming imperfect rules and storing occasional exceptions to those rules. Further-

Robert M. Nosofsky, Thomas J. Palmeri, and Stephen C. McKinley, Department of Psychology, Indiana University.

This work was supported by National Institute of Mental Health Grant PHS R01 MH48494-01 to Robert M. Nosofsky.

We would like to thank John Anderson, Larry Barsalou, Rob Goldstone, John Kruschke, Doug Medin, Rich Shiffrin, Thomas Ward, and an anonymous reviewer for their comments on earlier versions of this article and for helpful discussions.

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

more, according to such a view, the good fits of exemplar models may result from averaging over the responses of different subjects. This view of category learning is extremely interesting because it not only contrasts dramatically with current exemplar-trace theories but also establishes continuity with the early theorizing on hypothesis testing that once dominated the field of category learning.

If the general idea that averaged classification data represent mixtures of different rules and exceptions is to serve as a useful scientific construct, however, it is imperative to formalize explicit models of rule extraction and conduct rigorous empirical tests of such models. To date, no explicit, psychologically oriented rule-plus-exception models based on elementary processes of hypothesis testing have been tested on their ability to predict diverse phenomena in the modern classification literature. The purpose of the present research, therefore, was to initiate such a project by developing and providing rigorous tests of a formal rule-plus-exception model of classification learning.

In spirit, our model is almost the polar opposite of many of the currently popular exemplar models of classification. Learning is presumed to result from the extraction of simple logical rules using a process of hypothesis testing, with occasional exceptions to those rules also being stored. These exceptions are rarely complete exemplars; actual storage of complete exemplars occurs as a last resort. Thus, for any given subject, we imagine that category representations contain relatively little information, just a simple rule or two, supplemented by a few exceptions.

As we bring out in the General Discussion, our rule-plus-exception model (RULEX) also differs in important ways from a variety of rule-extraction models emanating from both the psychological and artificial intelligence literatures. In particular, rules are learned on an incremental, trial-by-trial basis by a process of hypothesis testing, and the system makes minimal memory requirements and information-processing demands on the part of the learner. Rather than designing a learning algorithm with goals such as speed of learning, efficiency, and optimality, the design of the present RULEX model was motivated primarily around considerations of psychological plausibility.

A key theme of our modeling is that subjects take themselves up by the bootstraps in solving classification problems, and the strategies that work for a given subject are often highly idiosyncratic. Large individual differences are expected to be observed in the particular rules that are extracted and the particular exceptions that are stored. As a result, extensive computer simulations are needed to use the RULEX model to predict classification data. Because of the complexities that are envisioned in the learning process, to get started on the theoretical investigation we need to restrict ourselves to a fairly simple domain of inquiry. Accordingly, we developed the present model to account for classification learning in domains in which the stimuli vary along separable binary-valued dimensions, in which there is a deterministic assignment of training exemplars to categories and in which there are two categories of exemplars to be learned. Fortunately, numerous experiments have been conducted in such domains, so there is a wealth of data for testing the model. Indeed, we demonstrate that the hypothesis-testing and simple rule-extraction processes that are formalized in our model can go a long way toward accounting for numerous fundamental categorization phenomena, including prototype effects, effects of specific exemplars, sensitivity to correlational information, the learning of linearly separable versus nonlinearly separable categories, selective attention effects, and the role of structural complexity in influencing speed of category learning.

Beyond accounting for these key phenomena, which serve as benchmarks for modern classification learning theories, we go on to demonstrate that the RULEX model may provide better predictions than extant alternative models of patterns of classification behavior observed at the individual subject level. With some notable exceptions (e.g., Ashby & Lee, 1991; Maddox & Ashby, 1993; Nosofsky, 1986), most current models of categorization have been tested on their ability to predict aggregate classification data. As is extremely well known, aggregate data may obscure patterns observed at the individual subject level. In the present research, we develop techniques that allow classification models to be compared on their ability to predict both aggregate data as well as distributions of individual subject behaviors from which the aggregate data are derived. We argue that these more fine-grained analytic techniques may be necessary to test adequately among the very powerful models of classification learning that have been developed in recent years.

In a nutshell, then, we suggest that complex and intricate patterns observed in aggregate classification data may not reflect complex and informationally rich category representations but rather averages computed over a number of different, fairly simple category representations, each of which has been learned through a simple process of hypothesis testing.

A Preliminary Example

Before presenting the model, which is formalized in a computer simulation, it is useful to provide an example of the process that is envisioned. Table 1 shows a category structure tested extensively by Medin et al. in numerous studies of classification learning (e.g., Medin & Schaffer, 1978). The stimuli vary along four binary-valued dimensions. There are five training exemplars in Category A and four in Category B; the remaining seven stimuli are transfer items that are not assigned by the experimenter to either category. In general, members of Category A tend to have a logical value of 1 on each of their dimensions, whereas members of Category B tend to have a logical value of 2. Importantly, however, no single-dimension rule is available for perfectly partitioning the exemplars into categories, and no

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

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

conjunctive rule is available either. The category structure is ill-defined in the sense that there are no singly necessary and jointly sufficient sets of features for determining category membership (Smith & Medin, 1981).

How might subjects solve such a categorization problem? Medin et al. (e.g., Medin, Dewey, & Murphy, 1983; Medin & Florian, 1992; Medin & Schaffer, 1978; Medin & Smith, 1981) and Nosofsky (1984, 1992) promoted an exemplar model of classification in which all exemplars of the categories are stored in memory and in which subjects compute summed similarities of items to these stored exemplars. This well-known context model of classification has provided excellent quantitative fits to numerous sets of classification learning data and far outperforms various prototype models that have been its main competitors (Estes, 1986b; Medin & Schaffer, 1978; Nosofsky, 1992).

In contrast, according to the RULEX model, the learning process might proceed as follows: One subject might notice that a value of 1 on Dimension 1 almost always signals Category A, whereas a value of 2 on Dimension 1 almost always signals Category B. So the subject starts by adopting a single-dimension rule along Dimension 1. As learning proceeds, the subject realizes that, although the rule is working fairly well, it is clearly not perfect. On encountering Stimulus 2111 from Category A and receiving corrective feedback, the subject then attempts to form an exception to the rule. For example, an exception might be formed that stimuli with the pattern 211* are also members of Category A. (In this notation, the asterisks denote wild cards that can take on any value.) The subject would quickly learn that such an exception does not work, however, because it matches Stimulus 2112, which is a member of Category B. Thus, this exception would be discarded, and a new exception would be formed on the next encounter of Stimulus 2111. Eventually, the subject might learn the rule that a value of 1 on Dimension 1 signals Category A, that a value of 2 signals Category B, but that the pattern 2*11 is an exception that belongs to Category A and the pattern 1*22 is an exception that belongs to Category B. Note that the classification problem is solved, even though no complete exemplars are stored in memory.

In contrast, a second subject might learn a completely different rule and set of exceptions, for example, that a value of 1 on Dimension 3 signals Category A, that a value of 2 on Dimension 3 signals Category B, and that the patterns 1*21 and 2*12 are exceptions to these rules. Other subjects might learn rules based on conjunctions of features, whereas still others might form rules and store exceptions that never allow them to solve the problem completely. Note that each individual subject is assumed to extract an extremely simple set of rules and exceptions. The aggregate data are predicted by mixing these idiosyncratic sets of rules and exceptions together.

Overview of the Formal Model

In this section, we provide an overview of RULEX. Our intent is to provide enough information to communicate the key conceptual underpinnings of the model. Some of the more technical details of the computer simulation are provided in the Appendix. We emphasize at the outset that we are not committed to many of the detailed processing assumptions in RULEX. The important issue is whether the general idea of mixing together idiosyncratic rules and exceptions provides a viable model of category learning and representation. Specific processing assumptions are needed to implement such a model in a computer simulation and to initiate the investigation. After our description of the formal computational model, we summarize the key conceptual aspects of the simulation that we consider to be the most important. Extensions of the model and relations between RULEX and other recently proposed rule-extraction models are considered in the General Discussion.

The general goal of the RULEX learning process is to form a "decision tree" in the sense of Hunt et al. (1966). The decision tree consists of a sequence of tests of the values of individual attributes in an object. The category into which an object is classified is determined by the outcome of this sequence of tests. For example, for the category structure shown in Table 1, one possible decision tree is illustrated in Figure 1. The first test asks whether the stimulus has a value of 1 or 2 on Dimension 1. If the stimulus has a value of 1, then it is tested for the exception pattern 1*22. A positive outcome on this latter test results in a Category B response, whereas a negative outcome results in a Category A response. Alternatively, if the stimulus has a value of 2 on Dimension 1, then it is tested for the exception pattern 2*11 and so forth (see Figure 1). Numerous classification algorithms in the artificial intelligence literature can be characterized as giving rise to such decision trees, but the tree-building process can vary considerably. Our aim in this research was to test a learning process that seemed psychologically plausible, learning such decision trees by developing and testing hypotheses on a trial-by-trial basis using induction over exemplars and making minimal memory requirements on the part of the learner. This set of characteristics distinguishes RULEX from other decision-tree building algorithms emanating from artificial intelligence (see General Discussion).

Computational Model

A schematic flow diagram of RULEX is shown in Figure 2. The first stage of learning involves a search for a perfect single-

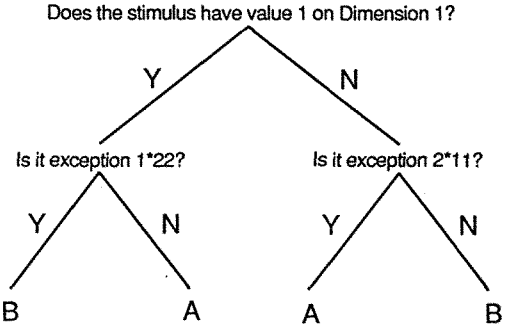


Figure 1. Schematic illustration of one possible decision tree for discriminating the members of Categories A and B in Medin and Schaffer's (1978) experimental paradigm (see Table 1). Y = yes; N = no. The terminal nodes of the decision tree indicate the category to which an item is assigned (A or B). Note that the tests for the exceptions (1*22 and 2*11) can themselves be broken down into a sequence of tests of values on the individual dimensions, thereby extending the decision tree. The simplified structure shown here is provided for conceptual clarity.

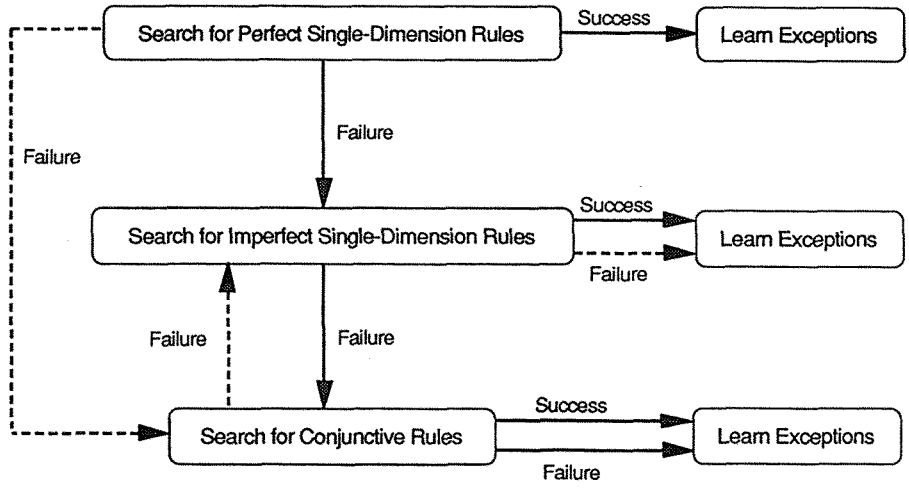


Figure 2. Schematic flow diagram of the sequence of hypothesis-testing stages in rule-plus-exception model of classification learning. The solid lines show the sequence that occurs with high probability, and the dotted lines show the sequence that occurs with lower probability.

dimension rule (cf. Levine, 1975; Trabasso & Bower, 1968). Each dimension is assigned a positive, real-valued weight that represents its intrinsic salience. In the first stage, an individual dimension is sampled with probability that is proportional to its weight. A single-dimension rule is then formed that is consistent with the exemplar and feedback that is received. For example, if Dimension 1 is sampled, the exemplar has a logical of value 1 on this dimension, and feedback for Category A is provided, then a subject would form the rule that exemplars with a value of 1 on Dimension 1 belong to Category A (1*** → A), whereas exemplars with a value of 2 on Dimension 1 belong to Category B (2*** → B). The subject continues to use a single-dimension rule as long as it works. If the rule fails to work perfectly, it is discarded, and a new dimension is sampled for testing a single-dimension rule. If no dimension yields a perfect rule, then a second stage of hypothesis testing begins.

In the second stage, subjects search for an imperfect, single-dimension rule, and, if this strategy fails, they search for a conjunctive rule in a third stage. (In the general version of the simulation, the ordering of these two strategies is probabilistic. However, in most of the simulations, we assume that the search for imperfect, single-dimension rules comes first, so we focus discussion on this special case.)

In searching for an imperfect, single-dimension rule, the subject samples dimensions by the same process as that used in Stage 1. A single-dimension rule is formed that is consistent with the presented exemplar and the feedback information. This rule is maintained for a minimum number of trials, termed the *lower test window.* Once the lower test window is reached, the single-dimension rule is maintained only as long as it exceeds a lax criterion; otherwise, it is discarded. For example, a subject might regard a single-dimension rule as tentatively acceptable as long as it correctly classifies 60% of the incoming exemplars. Once the *upper test window* is reached, the single-dimension rule is stored as a permanent rule as long as it exceeds a strict criterion; otherwise, it is discarded. For example, the imperfect rule might be acceptable as a permanent rule only if it correctly classifies at least 90% of the incoming exemplars. If an imperfect single-dimension rule is discarded, a new dimension is sampled and the search continues. If all of the dimensions are exhausted without an acceptable single-dimension rule being found, then a search begins for conjunctive rules.

In searching for conjunctive rules, pairs of dimensions are sampled with a probability that is proportional to the product of their weights. Each conjunctive rule is maintained until the lower test window is reached. Thereafter, the conjunctive rule is maintained only as long as it exceeds the lax criterion. If the conjunctive rule exceeds a strict criterion when the upper test window is reached, then it is stored as a permanent rule; otherwise, it is discarded, and a new pair of dimensions is sampled.

Once a permanent, single-dimension rule or conjunctive rule is formed, or once all single dimensions and pairs of dimensions are exhausted without a rule being formed, the subject then begins the exception-storage process. If an item is encountered that contradicts the rule or for which a rule does not apply, the subject probabilistically samples each of the item's dimensions. (The dimension that participates in the rule, however, is sampled with Probability 1.) For example, consider again the category structure in Table 1. Suppose that the subject had formed the rule that a value of 1 on Dimension 1 signals Category A. On encountering Stimulus 2111, Dimension 1 is sampled with Probability 1, and Dimensions 2 to 4 are individually sampled with some fixed sampling probability. If the pattern that is sampled is 211*, then the subject would attempt to remember that the exception 211* signals Category A. Two factors affect a subject's ability to remember an exception. First, the larger the number of dimensions that form the exception, the lower is the probability that it is successfully stored. Second, the larger the number of exceptions currently stored in memory, the lower is the probability that a new exception is successfully stored. These assumptions reflect the idea that the memory system is limited in capacity.

If an exception is stored that later produces an incorrect cat-

egorization decision, it is discarded from memory (although, at a subsequent point in the learning process, it could be sampled and stored again). Once stored in memory, exceptions are maintained as long as they continue to work.

Classification decisions proceed as follows. First, a check is made of all exceptions stored in memory. If an exception applies to the presented stimulus, it is used to make the decision. (For example, Exception 12** applies to Stimuli 1211, 1212, 1221, and 1222.) In the rare case that there are conflicting exceptions, a Category A response is made with probability equal to the proportion of exceptions that signal A. If no exceptions apply, then a check is made of any conjunctive rules that are stored. If no conjunctive rules are stored, then a check is made of any single-dimension rules that are stored. If no exceptions, conjunctive rules, or single-dimension rules apply, then a random guess is made. Note that classification decisions follow a specificity principle. The exceptions, which are the most specific rules, have highest priority in determining the classification decision (cf. Anderson, Kline, & Beasley, 1979).

Key Properties

Several conceptual points regarding the RULEX model should be emphasized. The most important general property of the model is that relatively little information about the original exemplars is stored in memory, at least for the purpose of making classification decisions.[1] In typical runs of the simulation, either a single-dimension rule or a conjunctive rule is stored, with just a couple of exceptions used to supplement the rule. The exceptions are rarely full exemplars. Indeed, because of the assumption of limited memory capacity, there is a bias in the model to store only subsets of the dimension values that compose individual exemplars. Note that, even during the learning process itself, there is essentially no memory for previously presented exemplars (cf. Trabasso & Bower, 1968). Subjects have memory only for previous hypotheses they have tested and for specific dimensional information relevant to the current hypothesis. Possibly, an improved learning model would result by allowing some short-term memory for previously presented exemplars (e.g., Fowler Williams, 1971; Levine, 1966). Such information could be used to make hypothesis selection and testing more efficient. The no-memory assumption is adopted here mainly for reasons of simplicity and to serve as a strong contrast against alternative exemplar-trace theories.

A second property of the model is that the particular rules and exceptions that are learned are highly idiosyncratic. Because of the probabilistic sampling of dimensions during the rule-formation process, the model predicts that, even after successfully solving a problem, individual subjects may vary greatly in the pattern of generalizations that they form when classifying transfer stimuli (cf. Medin, Wattenmaker, & Michalski, 1987; Nosofsky et al., 1989; Pavel et al., 1988). We also expect that subjects vary greatly in the criteria they apply for accepting imperfect rules, and this variability will also result in different patterns of generalization.

Another property of the model is that, early in learning, values on individual dimensions are expected to exert primary influence on classification decisions, whereas later in learning, when exceptions begin to be formed, combinations of values on dimensions start exerting influence. As will be seen, this relational coding aspect of storing occasional exceptions, together with the property that individual dimensions exert major influence, allows RULEX to account for a variety of important categorization phenomena, such as prototype effects, effects of specific exemplars, selective attention to dimensions, and sensitivity to correlational information.

Finally, we remark that, in its strong form, once the model has extracted a set of rules and exceptions, responding is predicted to be deterministic (cf. Ashby & Gott, 1988). In other words, a given subject who has solved a classification problem is predicted to always give the same categorization response for any given stimulus. Taken to an extreme, this prediction is implausible in that it imputes to the subject an automaton-like character without any of the fallibilities of human information processing. Even if the present RULEX model is correct in spirit, numerous factors exist that could introduce noise into the classification decision process, including lapses of attention, inadvertent button presses, and occasional trials in which the rules and exceptions are temporarily forgotten or misapplied. Thus, to develop a more realistic model, we assume that, on any given trial, there is some small probability that the subject gives the opposite response dictated by the extracted rules. Such an assumption has precedence and empirical support in the previous work of Levine (1966, 1975). This response-error parameter becomes important when we attempt to use RULEX to model detailed distributions of classification responses at the individual subject level, but for more standard applications we simply set it equal to zero.

Fitting the Model to Data

In this section, we apply RULEX to a variety of well-known experimental results in the categorization literature. Because RULEX is a simulation model and has a large number of available free parameters, the task of fitting the model to data is difficult. Our primary goal is not to obtain a best fit of the model to data but rather to show that even simple versions of the model with many of the parameters constrained can achieve reasonable accounts of major phenomena of interest.

The free parameters in the most general version of RULEX include the following (see Table 2 for a glossary): a set of weights to represent the salience of the individual dimensions that compose the stimuli; a general storage probability that determines the rate at which simple rules, conjunctions, and exceptions can be placed in memory; lax criteria for accepting tentative single-dimension and conjunctive rules; strict criteria for forming permanent single-dimension and conjunctive rules; lower test windows and upper test windows (numbers of trials) for testing these candidate rules; a branching parameter for determining whether to test conjunctive rules or imperfect, single-dimension rules first; a capacity-limit parameter that affects the rate at

[1] Later in this article, we argue that, although rules may be formed to classify exemplars, it is plausible that some subjects may also have memories for some of the exemplars that were presented. Although not used for classification, these stored exemplars could be used when subjects make recognition or typicality judgments or could be involved in implicit memory tasks such as perceptual identification.

Table 2
Glossary of Free Parameters in RULEX

Parameter	Description
w	Vector of dimension-salience weights (default: all weights set equal to 1.0)
pstor	General storage probability that affects the rate at which rules and exceptions can be added to memory and the probability that individual dimensions are sampled during the exception-formation process
lcrit	Lax criterion for tentatively accepting imperfect simple rules (default: lcrit = .55)
scrit	Strict criterion for forming a permanent, imperfect, single-dimension rule
ccrit	Strict criterion for forming a permanent, imperfect, conjunctive rule (default: ccrit = 1.0)
lwind	Lower test window (default: lwind = number of training items)
uwind	Upper test window (default: uwind = 2 × lwind)
branch	Probability that a subject tests imperfect, single-dimension rules before testing conjunctive rules (default: branch = 1.0)
capac	Capacity-limit parameter that affects the rate at which new exceptions can be added to memory (default: capac = 1.0)
rerr	Response-error parameter giving the probability that a subject makes the opposite response indicated by the extracted rules (default: rerr = 0.0)

Note. RULEX = rule-plus-exception model of classification learning.

which new exceptions can be added to memory; and a response-error parameter that allows for fallible application of extracted rules. Furthermore, each of these parameters is expected to vary across different subjects and might also be expected to vary within a single subject during the course of an experimental session. Thus, a fully adequate test of the model would require conducting simulations over probability distributions of these parameters.

To get started, however, we introduce numerous constraints on the parameter settings. For various applications, these constraints will be clear. For example, in some of the experiments, the assignment of physical dimensions to the abstract coding that defines the category structures is balanced over subjects. In these situations, the weights for representing differential intrinsic salience of the dimensions are not used. Certain arbitrary choices are made for various other parameters. For example, the lower test window is set equal to the number of training exemplars in the category, and the upper test window is set equal to twice the lower test window. The lax criterion for tentatively maintaining an imperfect rule is set at .55, which means that an imperfect rule is thrown out quickly only if it leads to nearly chance performance. In virtually all of the simulations, the probability of branching to single-dimension rules before testing conjunctive rules is set at 1. Also, in virtually all of the simulations, the parameter for representing a limited capacity on the number of exceptions that can be stored is not used because most of the experiments have a fairly limited number of training trials, leaving room for a relatively small number of exceptions to be stored in any case. Finally, except when fitting detailed individual subject distributions, the response-error parameter is not used. In sum, the applications of RULEX that are reported here make use of a reasonably small number of free parameters.

Applications to Empirical Phenomena

Medin and Schaffer (1978)

The first applications of RULEX are to the series of classification problems tested by Medin and Schaffer (1978) in their seminal article on the learning of ill-defined categories. A good initial illustration of RULEX is provided by considering Medin and Schaffer's Experiment 3. The category structure was the one shown in our informal example in Table 1. A set of Brunswik faces varying along the dimensions of eye height, eye separation, nose length, and mouth height were used as stimuli, and these physical dimensions instantiated the abstract category structure shown in the table. An unfortunate aspect of Medin and Schaffer's design (at least with respect to testing RULEX) is that each logical dimension was always instantiated by the same physical dimension. Thus, the dimension weights for representing differential intrinsic salience of the physical dimensions are needed for fitting the model. After an initial learning phase in which only the training stimuli were presented, Medin and Schaffer then conducted a test phase in which both the training and transfer stimuli were presented. The main goal is to use RULEX to predict the data observed during this test phase.

The free parameters that we used for fitting RULEX were the overall storage probability parameter (pstor), the strict criterion for single-dimension rules (scrit), the strict criterion for conjunctive rules (ccrit), and three freely varying dimension weights (w_1–w_4, in which the weights sum to 1). The general procedure for fitting the model was to conduct a discrete grid search over these parameters to find a combination of parameter values that produced a reasonable starting fit. Attempts were then made to fine-tune the fit by conducting a hill-climbing search, with the parameters obtained from the grid search used as starting values. Each run involved 5,000 Monte Carlo simulations of the rule-extraction process, with each simulation being conducted over a sequence of nearly 200 learning and transfer trials.[2] Thus, the model-fitting process was enormously time consuming. Furthermore, because of the intrinsic variability resulting from simulation methods, the parameter-search

[2] The precise learning criterion used by Medin and Schaffer (1978) varied across the different experiments. For simplicity, in the following simulations, we assume 16 blocks of learning trials, with each exemplar presented once in random order in each block.

process is made even more difficult, and the resulting fits are almost certainly local minima. Our goal, however, was not to find the absolute best fitting version of RULEX but simply to demonstrate that the model can provide a reasonable account of the data.

The predictions of RULEX are shown along with Medin and Schaffer's (1978) observed data in Table 3. The data are the probabilities with which each stimulus was classified in Category A during the transfer phase. Overall, the model gives a reasonable quantitative account of these data. The sum of squared deviations (*SSD*) between predicted and observed Category A response probabilities is .037, the root mean squared deviation (*RMSD*) is .048, and the model accounts for 98.0% of the variance. These fits are fairly close to those achieved by Medin and Schaffer's (1978) context model (*SSD* = .031).

It is interesting to understand how RULEX accounts for these data. The best fitting parameters are shown in Table 4. The most critical point is that the strict single-dimension rule criterion takes on a very low value (scrit = .55). Thus, according to the model, virtually all subjects adopted single-dimension rules in this experiment. Indeed, because searches for conjunctive rules never took place in these simulations, the value of the strict conjunctive rule criterion is arbitrary. As indicated by the values of the weight parameters, the dimensions of eye height (Dimension 1) and eye separation (Dimension 2) had the most intrinsic salience, followed by those of nose length (Dimension 3) and mouth height (Dimension 4). Of the 5,000 simulated subjects, roughly 60% adopted a single-dimension rule based on Dimension 1, almost 30% adopted a single-dimension rule based on Dimension 3, and the remaining 10% were split between Dimensions 2 and 4. Although Dimension 2 was more salient than Dimension 3, it is less diagnostic of category membership (see Table 3) and so is less likely to meet the criterion for forming an acceptable rule.

Table 3
Fit of RULEX Model to Medin and Schaffer's (1978) Experiment 3

Stimulus	Predicted *p*	Observed *p*
	Category A	
A1 1112	.950	.970
A2 1212	.974	.970
A3 1211	.997	.920
A4 1121	.867	.810
A5 2111	.734	.720
	Category B	
B1 1122	.391	.330
B2 2112	.210	.280
B3 2221	.026	.030
B4 2222	.001	.050
	Transfer	
T1 1221	.726	.720
T2 1222	.486	.560
T3 1111	.991	.980
T4 2212	.251	.230
T5 2121	.299	.270
T6 2211	.477	.390
T7 2122	.045	.090

Note. Entries are the predicted and observed probabilities with which each stimulus was classified in Category A during the test phase. RULEX = rule-plus-exception model of classification learning.

Table 4
Best Fitting Parameters for the RULEX Model Applied to Medin and Schaffer's (1978) Experiments

	Parameters						
Data set	pstor	scrit	ccrit	w_1	w_2	w_3	w_4
Experiment 3	.55	.55	1.00	.43	.33	.19	.05
Experiment 2	.65	.85	.55	.25	.06	.24	.44
Experiment 4	.58	.78	.55	.50	.25	.13	.13

Note. RULEX = rule-plus-exception model of classification learning; pstor = general storage probability; scrit = strict criterion for forming a permanent, imperfect, single-dimension rule; ccrit = strict criterion for forming a permanent, imperfect conjunctive rule; w_j = dimension j salience weight.

The general pattern of data in Table 3 is highly interpretable in terms of a model that posits that the majority of single-dimension rules were formed over Dimension 1 and that a sizable minority were formed over Dimension 3. As can be seen in the table, the training stimuli that satisfied these rules on both dimensions (A1, A2, and A3 of Category A and B3 and B4 of Category B) were classified with very high accuracy. The training stimuli that were exceptions to the Dimension 1 rule (A5 and B1) were classified least accurately, whereas the training stimuli that were exceptions to the Dimension 3 rule (A4 and B2) were next worst in accuracy. Apparently, after adoption of the single-dimension rules, many subjects failed to learn all of the exceptions that would be needed for perfect classification. In the present simulations, only about 20% of the subjects learned rules and exceptions that would allow for perfect performance. This figure accords fairly well with Medin and Schaffer's (1978) report that only 30% of their subjects reached a learning criterion of one errorless run through the set of nine training stimuli, because some of the correct responses could have been guesses or based on exceptions that were not fully stored in long-term memory.

The RULEX model predicts a couple of other critical phenomena in this data set. First, it predicts prototype effects. Transfer Stimulus T3 is the prototype of Category A, having a logical value of 1 on all of its dimensions. (Recall that members of Category A tend to have a logical value of 1 on each of their dimensions, whereas members of Category B tend to have a logical value of 2.) Although never presented during training, subjects classified T3 into Category A with extremely high probability. RULEX predicts this effect because essentially all of the single-dimension rules that could be adopted would lead T3 to be classified in Category A, and any exceptions that undo such rules would be temporary and rare.

Other successes for RULEX that are more subtle but of critical importance are its predictions of relative performance on Training Stimuli A1 and A2. Unfortunately, because of ceiling effects, the phenomenon is not evident in this set of observed

data. However, in virtually all of the experiments reported by Medin and his associates (e.g., Medin & Schaffer, 1978) that used this category structure, A2 was classified in Category A with higher probability than was A1. Note that A2 is less similar to the prototype of Category A than is A1. The stimuli match on Dimensions 1, 3, and 4, but A1 has a value of 1 on Dimension 2, whereas A2 has a value of 2 on this dimension. This effect of specific exemplars poses problems for a wide variety of models of classification, including prototype models and independent feature-frequency models (e.g., see Medin & Schaffer, 1978; Nosofsky, 1992).

As hinted at in Table 3 and documented more fully in some subsequent simulations, RULEX predicts this critical phenomenon. According to RULEX, A2 tends to have an advantage over A1 because various subconfigurations of dimension values that compose A1 also tend to appear in the contrast category. Thus, when exceptions are formed for classifying A1, they often need to be discarded because they lead to incorrect classifications of stimuli in the contrast category. For example, suppose that a single-dimension rule is formed over Dimension 4, with a logical value of 1 signaling Category A and a logical value of 2 signaling Category B (see Table 3). Stimuli A1 and A2 are exceptions to this rule. To classify A1, a subject might form the exception *112 → A. This exception, however, would then lead the subject to misclassify Stimulus B2 of Category B, so the exception would have to be discarded. More exceptions are available for A2 that do not lead to these misclassifications, so it tends to be easier to learn A2 than A1.

Another illustrative application of RULEX is provided in Table 5, which shows the fit of the model to Medin and Schaffer's (1978) Experiment 2. The same category structure was used as in Experiment 3. To instantiate the category structure, geometric forms were used as stimuli instead of the Brunswik faces in Experiment 3. Again, each logical dimension was always instantiated by the same physical dimension, so the dimension-salience weights are needed for fitting the model.

RULEX provides a fairly good account of the Experiment 2 data (*SSD* = .040, *RMSD* = .050, percentage of variance [%var] = 97.4). Again, the fit is almost certainly a local minimum but is already as good as that of the context model (*SSD* = .060). Interestingly, the best fitting version of RULEX extracted rules and exceptions that were quite different from those extracted in Experiment 3 (see the best fitting parameters in Table 4). Whereas in Experiment 3 the predominant type of rule was single dimensional, in Experiment 2 there was a strong bias to form conjunctive rules.[3] The value of the strict single-dimension rule criterion was quite high (.85), so permanent, single-dimension rules were rarely formed. According to the model, the most common strategies were to form conjunctive rules over Dimensions 1 and 4 and over Dimensions 3 and 4 and to supplement these rules by various exceptions. An example of a rule defined over Dimensions 1 and 4 was that the conjunction 1**1 signaled Category A, with all other items belonging to Category B, except that patterns 1*12 and 21*1 also belong to Category A. An example of a rule defined over Dimensions 3 and 4 was that the conjunction **11 signaled Category A, with all other items belonging to Category B, except that the patterns *121 and 1*12 also belong to Category A. Again, the rules and exceptions that were extracted were highly idiosyncratic, and there is no easy way to summarize them. It was again extremely common for the model to form rules and exceptions that did not provide a full solution to the problem. This result is consistent with the fairly high error rate displayed by Medin and Schaffer's (1978) subjects (see Table 5).

As was true for the previous data set, RULEX predicts the critical qualitative result that Stimulus A2 was classified with higher accuracy than was Stimulus A1. It predicts this result for two reasons. First, as discussed earlier, many of the subconfigurations of dimension values that compose A1 also tend to occur in the contrast category, so it is more difficult to learn exceptions for A1. Second, there are more conjunctive rules available for A2 than for A1 that do not conflict with members of the contrast category. Thus, there is a higher probability of forming a conjunctive rule that correctly classifies A2 than A1.

For completeness, Table 6 shows the fit of RULEX to Medin and Schaffer's (1978) Experiment 4. Note that a different category structure was used in this experiment. As before, logical values of 1 on each dimension tend to signal Category A, whereas logical values of 2 tend to signal Category B. The best fitting parameters are given in Table 4. RULEX's overall fit is not quite as good as before (*SSD* = .080, *RMSD* = .071, %var

Table 5
Fit of RULEX Model to Medin and Schaffer's (1978) Experiment 2

Stimulus	Predicted *p*	Observed *p*
	Category A	
A1 1112	.840	.780
A2 1212	.927	.880
A3 1211	.901	.810
A4 1121	.831	.880
A5 2111	.821	.810
	Category B	
B1 1122	.205	.160
B2 2112	.202	.160
B3 2221	.130	.120
B4 2222	.066	.030
	Transfer	
T1 1221	.601	.590
T2 1222	.376	.310
T3 1111	.830	.940
T4 2212	.364	.340
T5 2121	.492	.500
T6 2211	.572	.620
T7 2122	.139	.160

Note. Entries are the predicted and observed probabilities with which each stimulus was classified in Category A during the transfer phase. RULEX = rule-plus-exception model of classification learning.

[3] We do not have a good explanation as to why conjunctive rules may have been more prevalent in Medin and Schaffer's (1978) Experiment 2 than in their Experiment 3. The parameter values needed to fit the context model to these data also differ markedly across these two experiments, and there has been no principled explanation of this finding either.

Table 6
Fit of RULEX Model to Medin and Schaffer's (1978) Experiment 4

Stimulus	Predicted *p*	Observed *p*
	Category A	
A1 2112	.72	.80
A2 1112	.91	.78
A3 2111	.79	.86
A4 1212	.78	.83
A5 1121	.71	.72
A6 1211	.74	.80
	Category B	
B1 1221	.25	.30
B2 1222	.21	.36
B3 2212	.20	.27
B4 2211	.21	.22
B5 2122	.27	.31
	Transfer	
T1 1111	.86	.89
T2 2121	.56	.56
T3 2222	.20	.22
T4 1122	.53	.62
T5 2221	.30	.36

Note. Entries are the predicted and observed probabilities with which each stimulus was classified in Category A during the test phase. RULEX = rule-plus-exception model of classification learning.

= 92.0) but is still only slightly worse than that of the context model (*SSD* = .078). The model correctly predicts the prototype enhancement effects observed in the experiment (i.e., the high classification accuracies for the transfer patterns 1111 and 2222). The rules and exceptions extracted by the model were again highly idiosyncratic and involved a complex mixture of single-dimension rules, conjunctive rules, and exceptions.

In summary, RULEX provides reasonable quantitative accounts of the classification transfer data observed in Medin and Schaffer's (1978) experiments. It also predicts some of the salient qualitative results of those experiments, such as prototype effects and effects of specific exemplars. These effects are predicted despite the fact that the learning process in RULEX is quite different from the one assumed in the context model. In the present simulations of RULEX, complete exemplars were rarely stored, and similarity comparisons to exemplars are not involved in the classification decision process. Instead, the averaged classification data are conceptualized as mixtures of idiosyncratic rules and exceptions.

Evolution of Generalizations as a Function of Learning

Because RULEX assumes that people first test single-dimensional rules and then later form higher dimensional rules (i.e., store exceptions), it predicts changing patterns of generalization as a function of learning. In general, whereas dimensions that are individually diagnostic should exert primary influence on generalization early in learning, combinations of dimensions that are diagnostic can show more of an influence later in learning. Medin, Altom, Edelson, and Freko (1982) reported an important experiment that tested the extent to which people attend to individual dimensions as opposed to combinations of dimensions. The category structure they designed is useful for testing RULEX's predictions about changing patterns of generalization.

Medin et al.'s (1982) category structure is shown in Table 7. The stimuli vary along four binary-valued dimensions. Stimuli A1 to A4 are training exemplars of Category A, Stimuli B1 to B4 are training exemplars of Category B, and Stimuli T1 to T8 are new, unassigned transfer stimuli. The values on Dimensions 3 and 4 are perfectly correlated in the training set, such that their combination serves as a perfect predictor of category membership. However, neither of these dimensions is individually diagnostic; each of the dimension values is associated with the alternative categories 50% of the time. In contrast, Dimensions 1 and 2 are individually diagnostic. A value of 1 on each dimension predicts Category A 75% of the time, and a value of 2 on each dimension predicts Category B 75% of the time.

This category structure therefore pits individual-dimension diagnosticity against correlated-dimensions diagnosticity. Furthermore, the manner in which people classify the transfer stimuli can provide clues about which dimensions are exerting influence on people's classification decisions. To the extent that the correlated dimensions are exerting primary influence, peo-

Table 7
Observed Data From Medin, Altom, Edelson, and Freko's (1982) Design That Pitted Correlated Features Against Individually Diagnostic Ones

	Data set			
Stimulus	1	2	3	4
	Category A			
A1 1111	.88	.99	.96	.64
A2 2111	.89	.98	.93	.64
A3 1122	.73	.99	1.00	.66
A4 1222	.77	.95	.96	.55
	Category B			
B1 1212	.12	.01	.02	.57
B2 2212	.17	.01	.00	.43
B3 2121	.25	.01	.05	.46
B4 2221	.33	.00	.00	.34
	New transfer			
T1 2222	.53	.58	.66	.46
T2 2211	.53	.56	.64	.41
T3 2122	.75	.71	.64	.52
T4 1211	.67	.70	.66	.50
T5 1112	.45	.46	.36	.73
T6 1121	.38	.45	.36	.59
T7 2112	.36	.40	.27	.39
T8 1221	.28	.26	.30	.46

Note. 1 = Medin, Altom, Edelson, and Freko (1982); 2 = Pavel, Gluck, and Henkle (1988); 3 = McKinley and Nosofsky (1993), final transfer block; 4 = McKinley and Nosofsky (1993), first transfer block. Entries are the probabilities with which each stimulus was classified in Category A during the transfer phase.

ple should tend to classify transfer stimuli into Category A when the values on Dimensions 3 and 4 agree but classify transfer stimuli into Category B when these values disagree. On the other hand, to the extent that diagnosticity of individual dimensions is important, people should tend to classify transfer stimuli into Category A when they have a value of 1 on Dimensions 1 and 2 and into Category B otherwise.

In Medin et al.'s (1982) experiment, subjects freely inspected the eight training exemplars during a 10-min period to learn their category assignments. After this period, they then classified each of the training exemplars and the transfer items. The results are shown in Table 7 (column 1) in terms of the probability with which each pattern was classified in Category A. The data indicate that the subjects were sensitive to values on the correlated dimensions, because novel patterns that preserved the correlation (T1–T4) were classified primarily in Category A, whereas novel patterns that broke the correlation (T5–T8) were classified primarily in Category B. Some sensitivity to individual dimension diagnosticity is also evident, however, because patterns T3 and T4 were classified in Category A more often than patterns T1 and T2 and patterns T5 and T6 were classified in Category A more often than patterns T7 and T8.

Two follow-ups of the Medin et al. (1982) experiment have been conducted. Pavel et al. (1988) used Medin et al.'s category structure, but, instead of using a free-inspection paradigm, subjects learned to classify the training patterns in a supervised learning paradigm. Like Medin et al., they then conducted a transfer test in which both the training patterns and novel patterns were presented. The averaged classification probabilities for those subjects meeting a learning criterion are shown along with Medin et al.'s data in Table 7 (column 2). Although Pavel et al.'s criterion subjects learned to classify the training patterns with higher accuracy than Medin et al.'s subjects, the pattern of transfer data on the novel patterns was essentially identical in the two studies.

McKinley and Nosofsky (1993) conducted another replication and extension of this learning paradigm. The only important difference from Pavel et al.'s (1988) experiment was that, in addition to conducting a transfer phase at the end of learning, McKinley and Nosofsky inserted transfer phases at intermediate points during the learning sequence. As reported in Table 7 (column 3), the data obtained in the final transfer phase after the completion of learning show the same pattern as was observed for Medin et al.'s (1982) and Pavel et al.'s subjects.

However, the data obtained during the first transfer phase after the initial block of learning show a much different pattern (column 4). In general, the dimensions that are individually diagnostic appear to exert primary influence. (Pavel et al., 1988, reported a similar result for their learning data.) For example, consider Transfer Stimuli T1, T2, T5, and T6. These transfer stimuli are "conflicting items" in which the correlated dimensions and the individually diagnostic dimensions are in direct competition. For T1 and T2, the correlated dimensions point to Category A, but the individually diagnostic dimensions both point to Category B. The reverse holds for T5 and T6. Averaging over these four stimuli, subjects chose the category indicated by the correlated dimensions with a probability of .387 and chose the category indicated by the individually diagnostic dimensions with a probability of .613. This pattern contrasts dramatically with that observed during the final block of transfer, in which the correlated-dimensions choice was made with an average probability of .642 and the individual-dimensions choice was made with an average probability of .358.

In summary, the transfer data obtained by Medin et al. (1982), Pavel et al. (1988), and McKinley and Nosofsky (1993) provide evidence that, by the end of learning, the correlated dimensions exert a major influence on subjects' classification decisions, with some residual influence from the individually diagnostic dimensions. Early in learning, however, it appears that the individually diagnostic dimensions exert primary influence.

RULEX accounts for this pattern of results. Without an extensive parameter search, we were able to find reasonable settings on the parameters that allowed RULEX to achieve a good quantitative description of McKinley and Nosofsky's (1993) data. Because McKinley and Nosofsky randomized the assignment of physical dimensions to the logical category structure, the dimension-salience weights were not used (i.e., they were all set equal to 1.0). The general storage parameter was set at pstor = .6, and the strict criterion for conjunctive rules was set at ccrit = 1.00. To simulate the data, we assumed that there was variability in the value of the strict criterion for accepting imperfect single-dimension rules. This value varied uniformly from .7 to .9. This distribution of values represents the idea that some subjects may set a fairly lenient criterion for accepting single-dimension rules, and other subjects may set a strict one. Finally, preliminary simulations suggested that the default assumption of setting the upper test window at twice the size of the training set needed to be modified (learning proceeded too slowly). Instead, the value of the upper test window was set equal to the value of the lower test window (8).

The predictions of RULEX, based on 5,000 simulated subjects, are shown in Table 8. Overall, RULEX provides an excellent account of the averaged classification data. Consider first the transfer data from the final block. Averaged over the patterns T1, T2, T5, and T6 (i.e., the conflicting items), RULEX predicts that the correlated-dimension choice is made with a probability of .600 compared with the observed value of .642. For patterns T3, T4, T7, and T8 (i.e., the nonconflicting items), RULEX predicts that the correlated-dimensions choice is made with a probability of .706 compared with the observed value of .682. Thus, at the final block of transfer, RULEX predicts the primary influence of the correlated dimensions and the residual influence of the individually diagnostic ones. RULEX also predicts some subtle patterns involving the training exemplars. Note that, in the observed data, the overall percentage correct for the nonconflicting training exemplars (A1, A3, B2, and B4) was .989, whereas for the conflicting training exemplars (A2, A4, B1, and B3), overall percentage correct was .955. These values are well predicted by RULEX (predicted values of .995 and .933, respectively).

In addition to predicting the pattern of transfer data at the final block of learning, RULEX predicts the pattern of transfer data after the initial block of learning. Overall percentage correct on the training patterns is predicted to be .563, which is reasonably close to the observed value of .585. For the novel patterns in which the correlated dimensions are in competition with the individually diagnostic dimensions (T1, T2, T5, and T6), RULEX predicts that the correlated-dimensions choice is

Table 8
Fit of RULEX Model to McKinley and Nosofsky's (1993) Replication and Extension of Medin, Altom, Edelson, and Freko's (1982) Study

	Final block		First block	
Stimulus	Predicted *p*	Observed *p*	Predicted *p*	Observed *p*
		Category A		
A1 1111	.994	.96	.622	.64
A2 2111	.934	.93	.511	.64
A3 1122	.993	1.00	.612	.66
A4 1222	.924	.96	.489	.55
		Category B		
B1 1212	.064	.02	.491	.57
B2 2212	.004	.00	.378	.43
B3 2121	.062	.05	.499	.46
B4 2221	.004	.00	.382	.34
		Transfer		
T1 2222	.591	.66	.378	.46
T2 2211	.592	.64	.381	.41
T3 2122	.705	.64	.489	.52
T4 1211	.701	.66	.503	.50
T5 1112	.406	.36	.612	.73
T6 1121	.399	.36	.616	.59
T7 2112	.293	.27	.491	.39
T8 1221	.288	.30	.496	.46

Note. RULEX = rule-plus-exception model of classification learning.

made with probability .383. This prediction is reasonably close to the observed value of .387. For the novel patterns in which ambiguous information is provided by the individually diagnostic dimensions (T3, T4, T7, and T8), RULEX predicts that the correlated-dimensions choice is made with probability .501, which is fairly close to the observed value of .546. In summary, RULEX provides reasonable quantitative predictions of the averaged classification data observed both at the final block of transfer and after the initial block of transfer and predicts how the pattern of generalization evolves as a function of learning.

According to RULEX, averaged classification data often reflect a mixture of highly disparate patterns of responding at the individual subject level. Thus, it is critical to understand the patterns of generalization exhibited by individual subjects. In the study by Pavel et al. (1988), a generalization profile was defined for each subject. The profile was given by the sequence of Category A and Category B responses made for each of the novel transfer stimuli, T1 to T8. For example, the profile AAAABBBB reflects a subject that classified Transfer Stimuli T1 to T4 in Category A and T5 to T8 in Category B. Pavel et al. reported a wide variety of generalization profiles for their individual subjects, although an actual frequency distribution of such profiles was not reported. McKinley and Nosofsky (1993) conducted a similar analysis. They too observed a wide variety of generalization profiles.

The critical question is whether RULEX (and other models) can predict the distribution of generalizations that was observed. Given the parameter settings discussed earlier, RULEX predicts that 51.9% of the subjects should have adopted the generalization profile AAAABBBB. This profile is the one predicted for subjects who adopt the pure correlated-dimension rule: **11 → A, **22 → A, **12 → B, **21 → B. In McKinley and Nosofsky's (1993) experiment, the generalization profile AAAABBBB was actually observed for 47.7% of the subjects (21 of 44), which is quite close to the predicted value. RULEX further predicts that 43.8% of the subjects should have adopted 17 other moderate-frequency generalization profiles (each profile is predicted for at least 1% of the subjects). Most of these profiles correspond to a wide variety of single-dimension-plus-exception rules. For example, profile BBBAAABA is consistent with the rule 1*** → A, 2*** → B, with the memorized exceptions 2111 → A and 1212 → B. In McKinley and Nosofsky's (1993) experiment, 29.5% of the subjects actually adopted one of these profiles, which is somewhat less than predicted. Finally, RULEX predicts that only 4.3% of the subjects should have adopted one of the remaining possible 238 low-frequency profiles. In actuality, 22.7% (10 of 44) of McKinley and Nosofsky's subjects did so (although none of these 238 profiles was adopted by more than 1 subject). By setting the response-error (rerr) parameter to a small value (e.g., rerr = .02), the fit to the averaged transfer data is essentially the same as before, but better predictions of the distribution of generalization profiles can be achieved: correlated-dimension profile, 44.1% predicted, 47.7% observed; single-dimension-plus-exception profiles, 40.0% predicted, 29.5% observed; other profiles, 16.0% predicted, 22.7% observed.

In summary, in addition to accounting for the averaged classification transfer data observed at both early and late stages of learning, RULEX can provide at least a fair account of the pattern of generalizations observed for individual subjects. This goal of predicting distributions of responses at the individual subject level is quite ambitious but may be necessary to test adequately among the very powerful models of classification learning that have been developed in recent years. We renew this method of testing models in a subsequent section of our article.

Linearly Separable Versus Nonlinearly Separable Categories

An important issue in categorization concerns the role of linear separability in classification learning. Two categories are linearly separable if they can be partitioned by a linear discriminant function. In a two-dimensional space, this means that a straight line exists such that all members of Category A fall to one side of the line and all members of Category B fall to the other side of the line. Another way of thinking about linear separability is that one should be able to sum evidence along each individual dimension of an object and be able to determine its category membership depending on whether this summed evidence exceeds a criterion value.

The construct of linear separability is important because numerous models of classification predict that linearly separable categories should be easier to learn than nonlinearly separable ones. Such models include simple prototype models (Reed, 1972), independent feature-frequency models (Estes, 1986a, 1986b), and the fuzzy-logical model of perception (Massaro & Friedman, 1990). Although standard back-propagation models

(Rumelhart, Hinton, & Williams, 1986) are able to learn nonlinearly separable categories, even these models tend to predict that linearly separable categories should be easier to learn than nonlinearly separable ones (Gluck, 1991; Kruschke, 1993).

Medin and Schwanenflugel (1981) tested this prediction in a series of four experiments that used stimuli varying on binary-valued dimensions. In each experiment, they designed two category structures: one that was linearly separable and one that was not. In contrast to the predictions of numerous models, in none of the experiments were the linearly separable categories learned with fewer errors than the nonlinearly separable ones. (This result does not imply, of course, that linearly separable categories will never be easier to learn. It implies only that it is possible to design structures in which the predictions of the extant models are not obtained.)

An example of one of the designs tested by Medin and Schwanenflugel (1981) is shown in Table 9. The linearly separable categories are shown in the top part of the table, and the nonlinearly separable categories are shown in the bottom part of the table. The results of their experiment are shown in the top panel of Figure 3. The figure plots the average probabilities of errors for the linearly separable and nonlinearly separable categories as a function of blocks of learning. As can be seen, the nonlinearly separable categories were learned with fewer errors than were the linearly separable ones.

The predictions of RULEX are shown in the bottom panel of Figure 3. In fitting the model, we set the value of the strict single-dimension rule criterion at a moderate value (.65) and set the upper test window equal to the lower test window (which was held at its default value). We then adjusted the value of the general storage probability parameter until we achieved a reasonable fit to the data. (All other parameters were held at their default values.) As can be seen in Figure 3, RULEX correctly predicts that the nonlinearly separable categories are learned with fewer errors than the linearly separable ones and provides a good quantitative match to the data.

How does RULEX capture this advantage of the nonlinearly separable categories? As can be seen in Table 9, for both the linearly separable and nonlinearly separable categories, the value of 1 on each dimension points to Category A two thirds of the time, and the value of 2 points to Category B two thirds of the time. Thus, subjects would be equally fast at extracting

Table 9
Structure of Medin and Schwanenflugel's (1981, Experiment 3) Categorization Problems

Category A	Category B
Linearly separable	
A1 2111	B1 1222
A2 1112	B2 2221
A3 1221	B3 2112
Nonlinearly separable	
A1 1122	B1 2222
A2 2211	B2 2121
A3 1111	B3 1212

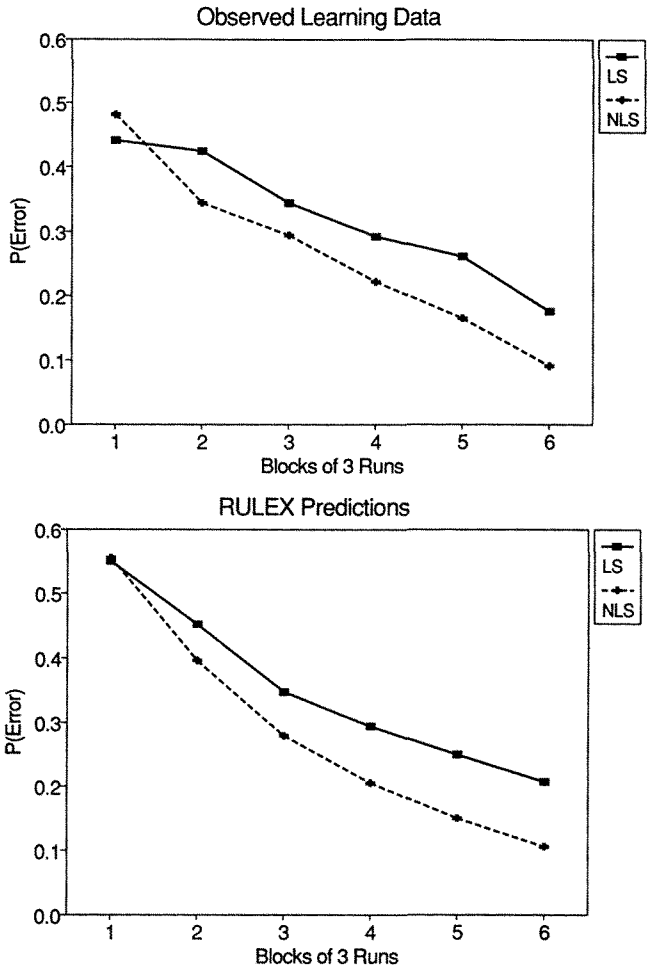


Figure 3. Top panel: Observed learning data from Medin and Schwanenflugel's (1981) Experiment 3 (solid lines are for the linearly separable [LS] categories and dashed lines are for the nonlinearly separable [NLS] categories). Bottom panel: Predicted learning data from the rule-plus-exception (RULEX) model of classification learning.

single-dimension rules for both of these structures. However, once the search for exceptions to those rules begins, the nonlinearly separable structure has a clear advantage. Consider all pairwise combinations of dimension values in the two structures that might be used for forming exceptions. In the nonlinearly separable case, there are 10 pairwise combinations of the dimensions in the exemplars of each category that are unique to that category. For example, in Exemplar A1, the combination 1*2* occurs in Category A but not in Category B. By contrast, in the linearly separable case, it turns out that there are only five pairwise combinations of the dimensions in the exemplars of each category that are unique to that category.

Because the linearly separable categories have fewer unique pairwise combinations, it is more difficult to locate exceptions that allow the problem to be solved. For example, for the linearly separable categories, suppose that a subject extracted the single-dimension rule 1*** → A and then formed the exception 2*1* → A when encountering A1. This exception would then

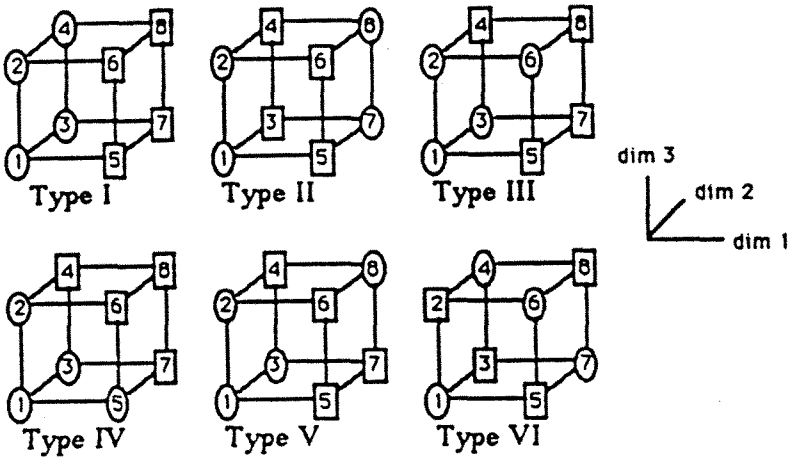
be discarded when the subject later encountered B3. In general, it takes RULEX longer to find workable exceptions for the linearly separable categories than for the nonlinearly separable ones, so the nonlinearly separable categories are learned more quickly.

Category Structure and the Speed of Learning Classifications

In the previous section, we illustrated RULEX's ability to predict that certain nonlinearly separable structures are learned more quickly than certain linearly separable ones. A more intricate set of learning data was recently reported by Nosofsky, Gluck, Palmeri, McKinley, and Glauthier (in press), who conducted replications and extensions of the classic study of Shepard, Hovland, and Jenkins (1961). This set of data allows for still more rigorous tests of alternative models of classification learning (cf. Anderson, 1991; Estes, in press; Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, 1984).

In Shepard et al.'s (1961) study, subjects were tested on six types of classification problems. In each problem, there were eight stimuli composed of three binary-valued dimensions. Four of the stimuli belonged to one category and the other four stimuli to a second category. The six problem types that result from these constraints are illustrated in Figure 4, with the stimuli represented as the vertices of a cube. Assignment of stimuli to each category is indicated by oval (Category A) or rectangular (Category B) vertices. Each face of the cube represents a value along one of the binary-valued dimensions. For ease of discussion, we imagine that the dimensions correspond to shape (square vs. triangle), color (black vs. white), and size (large vs. small), as is illustrated in the bottom part of the figure. Any assignment of stimuli to categories, with four stimuli in each category, can be rotated or reflected into one of the cubes shown in the figure.

The simplest category structure is the Type I problem. Here, information about only one dimension is necessary to solve the problem (shape in the example in Figure 4). For Type II, exactly two dimensions are relevant. In the Figure 4 example, black squares and white triangles are assigned to Category A, whereas white squares and black triangles are assigned to Category B. Information about size is irrelevant to solving the problem. Type VI is the most complex category structure, with all three dimensions being equally relevant. Stating a logical rule for



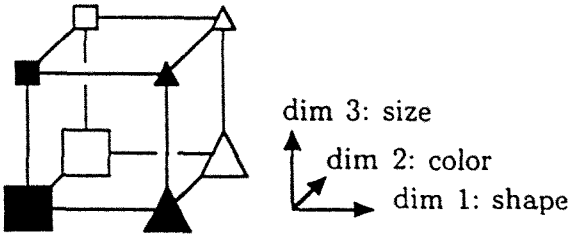


Figure 4. Schematic illustration of the six types of categorization problems tested by Shepard, Hovland, and Jenkins (1961). dim = dimension. From "Learning and Memorization of Classifications" by R. N. Shepard, C. L. Hovland, and H. M. Jenkins, 1961, *Psychological Monographs, 75* (13, Whole No. 517), p. 4.

Type VI in terms of the values on each of the component dimensions amounts to enumerating the stimuli in each of the categories. Finally, Types III, IV, and V are intermediate in structural complexity between Types II and VI. All three dimensions are relevant but to differing extents. One way of describing these problems is as single-dimension-plus-exception structures. For example, in Type V, squares are assigned to Category A and triangles to Category B, except that the small, white square is switched with the small, white triangle. The nature of the exceptions varies across the Type III, IV, and V problems, a point to which we return later.

In Shepard et al.'s (1961) original study, the Type I problem was learned with the fewest errors followed by Type II, Types III, IV, and V (which were about equal in difficulty), and finally Type VI. Nosofsky et al. (in press) conducted an extensive replication of Shepard et al.'s study that involved far more subjects, and they collected sufficient data to obtain learning curves. Thus, in addition to providing information about the total errors for each problem type, their study provided information about the actual time course of classification learning. Their results are reproduced in Figure 5, which plots the average probability of errors in each of the problems as a function of blocks of learning. (Each of the eight stimuli was presented twice in random order in each block of 16 trials.) The results corroborate Shepard et al.'s original findings and provide more detailed information about the actual learning process.

The predictions of RULEX are shown next to the observed data in Figure 5. As can be seen, the fit of the model to data is excellent (*SSD* = .077, *RMSD* = .028, %var = 94.4). The free parameters used for fitting RULEX were pstor = .80, scrit = .75, uwind = 4, branch = .10, and capac = .40. All other parameters were set at their default values, so that a five-parameter model was used to predict the 96 freely varying data points. (Because Nosofsky et al., in press, balanced the assignment of physical dimensions to the logical category structure, the dimension salience weights were not needed for fitting the data.) The main role of the capacity-limiting parameter was to reduce the speed of learning of the Type VI problem. When the capacity-limiting parameter is not used, RULEX still predicts that the Type VI problem is learned more slowly than the other types but fails to predict the magnitude of the effect. Likewise, when the branching parameter is not used, RULEX still predicts that the Type II problem is learned more quickly than Types III, IV, and V. By allowing for quick branches for testing conjunctive rules, however, RULEX achieves better quantitative fits to the data.

The ability of RULEX to predict these learning data falls naturally out of the rule-search and exception-learning processes. The Type I problem is learned most rapidly because subjects quickly extract the perfect, single-dimension rule that defines its structure. The Type II problem is learned next most rapidly because the search for conjunctive rules also occurs early in the learning sequence, and a perfect set of conjunctive rules is available for solving the Type II problem. Although a search for imperfect, single-dimension rules occurs early for the Type III, IV, and V problems, the rate of learning for these problems is slowed for a couple of reasons. First, in many cases, the single-dimension rule does not exceed the criterion set by the subject, so it is discarded. Second, even for those cases in which a single-dimension rule is formed, an additional stage is required for subjects to learn the exceptions, and this stage can be quite time consuming. Finally, Type VI lags considerably behind the other problems because the only way for RULEX to solve this problem is to store all of the complete exemplars in memory.

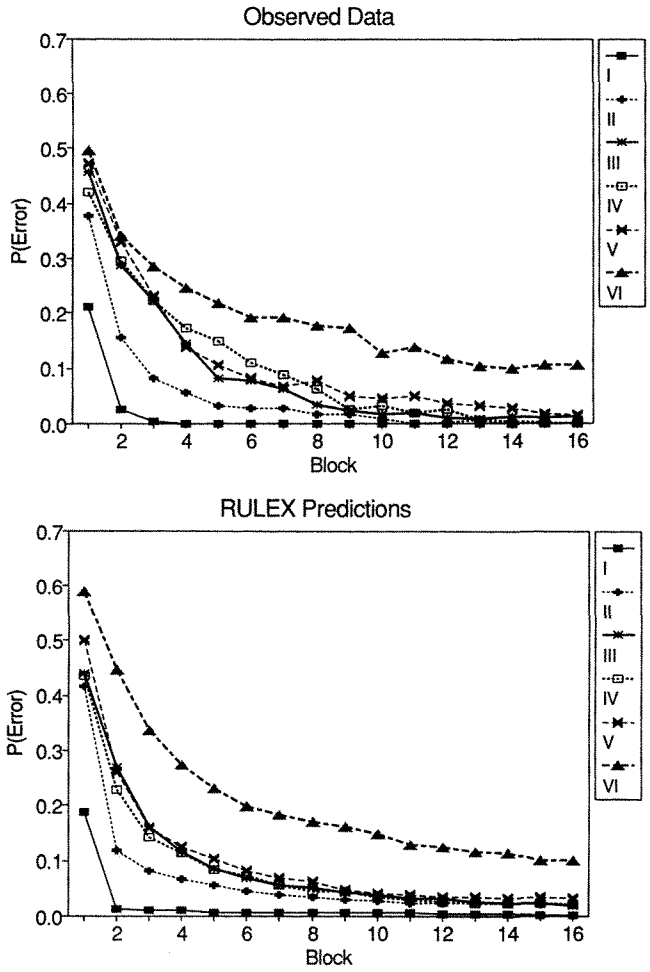


Figure 5. Top panel: Observed learning data from Nosofsky, Gluck, Palmeri, McKinley, and Glauthier's (in press) replication of Shepard, Hovland, and Jenkins's (1961) experiment. Bottom panel: Predicted learning data from rule-plus-exception (RULEX) model of classification learning. From "Comparing Models of Rule-Based Classification Learning: A Replication of Shepard, Hovland, and Jenkins (1961)" by R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley, and P. T. Glauthier, in press, *Memory & Cognition.* Copyright by the Psychonomic Society.

A more fine-grained breakdown of the experimental data is shown in Figures 6 to 8, which plot learning curves for individual types of items in Problems III, IV, and V. (In Problems I, II, and VI, all items play the same structural role and are logically equivalent.) In Problem IV, for example, Items 1 and 8 can be described as "central" members of their categories and Items 2 to 7 as "peripheral" members (see Figure 4). (One way of seeing why Items 1 and 8 are central is to notice that, regardless of which dimension is selected, these items will never be excep-

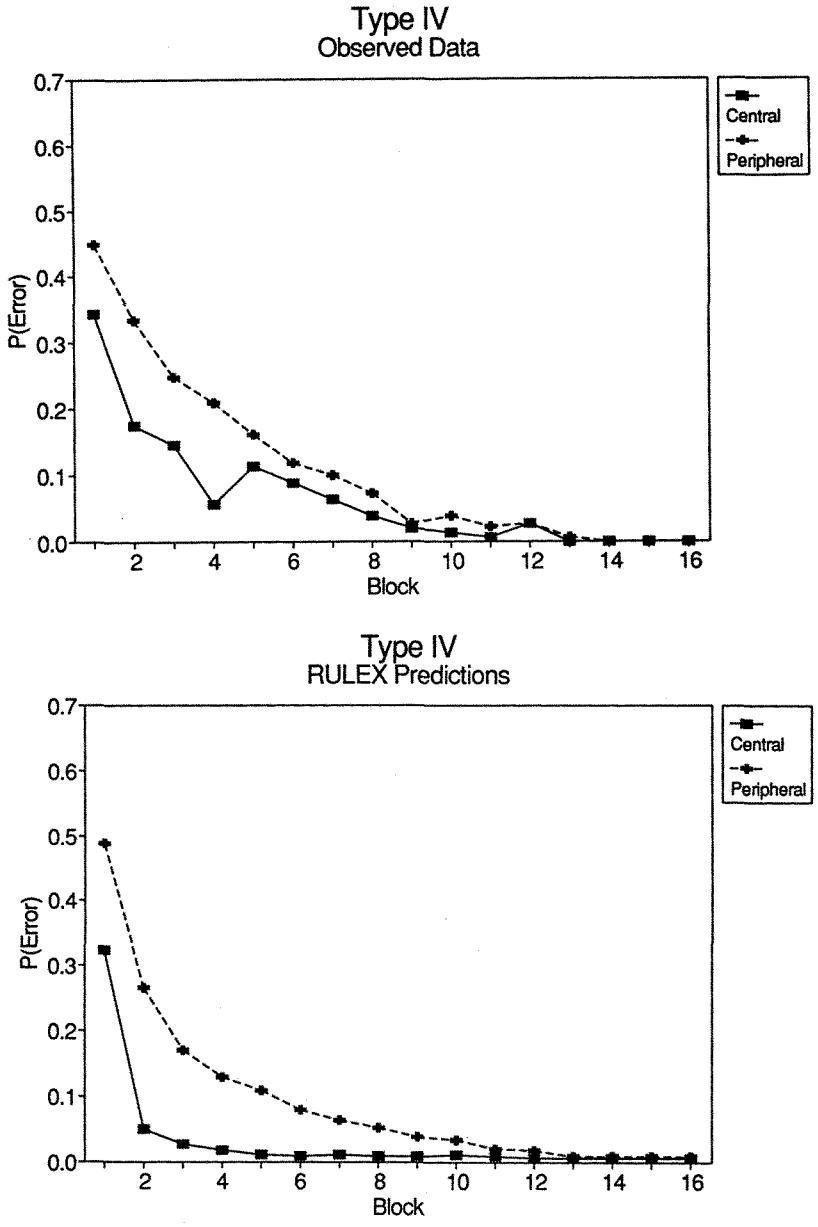


Figure 6. Observed and predicted learning data for the Type IV problem. RULEX = rule-plus-exception model of classification learning.

tions to the single-dimension rule.) Furthermore, the central members are logically equivalent to one another, and all peripheral members are logically equivalent to one another in the sense that they can be interchanged for one another by simple reassignment of dimensions or category labels. Thus, the data are averaged over the stimuli that define these item types. As shown in Figure 6, the central members of the Type IV problem were learned with fewer errors than were the peripheral members. In the Type III problem, four stimuli are central members (1, 2, 7, and 8) and four stimuli are peripheral members (3, 4, 5, and 6). Again, the central members were learned with fewer errors than were the peripheral members (see Figure 7). In the Type V problem, there are three item types, which we describe as central (1 and 5), peripheral (2, 3, 6, and 7), and exceptions (4 and 8). (Items 4 and 8 are exceptions to the only single-dimension rule that is available for the Type V problem.) As shown in Figure 8, the central members were learned with fewer errors than were the peripheral members, and the exceptions had the most errors.

The predictions of the RULEX model for the individual item types are shown next to the observed data in Figures 6 to 8. (The parameters were held fixed at those values given earlier.) At least at a qualitative level, RULEX correctly predicts all of the trends noted previously here. The central members have

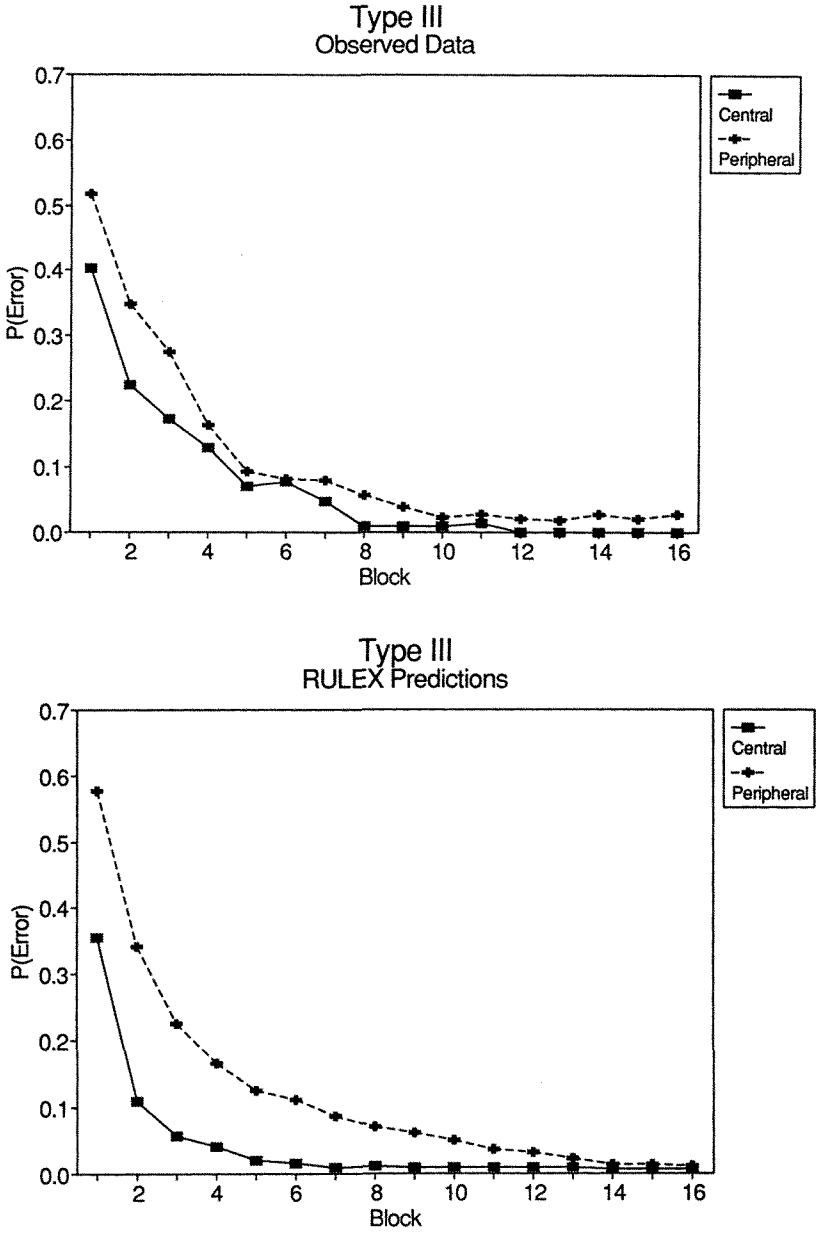


Figure 7. Observed and predicted learning data for the Type III problem. RULEX = rule-plus-exception model of classification learning.

fewer errors than the peripheral members because they are more likely to be correctly classified by the simple rules that are extracted early in the learning process. It is more likely that exceptions are needed to classify the peripheral members correctly, and this stage occurs later in the learning sequence. The exceptions in the Type V problem have the most errors because they are always learned during the exception-storage stage.

Predicting Distributions of Generalizations

Earlier in our article, we introduced the idea of testing classification models on their ability to predict distributions of classification responses at the individual subject level. The idea was to define the pattern of generalization exhibited by each individual subject and then use the models to predict the distribution of generalizations.

It is worth emphasizing the potential importance of this approach. An alternative idea is to try to fit a model to each individual subject's data. The problem with such a method, however, is that RULEX views classification learning as an inherently stochastic process. One cannot predict ahead of time which particular dimension a subject might sample when testing a rule or which dimensions might enter into the exceptions that are formed. Thus, fitting any given individual subject's data

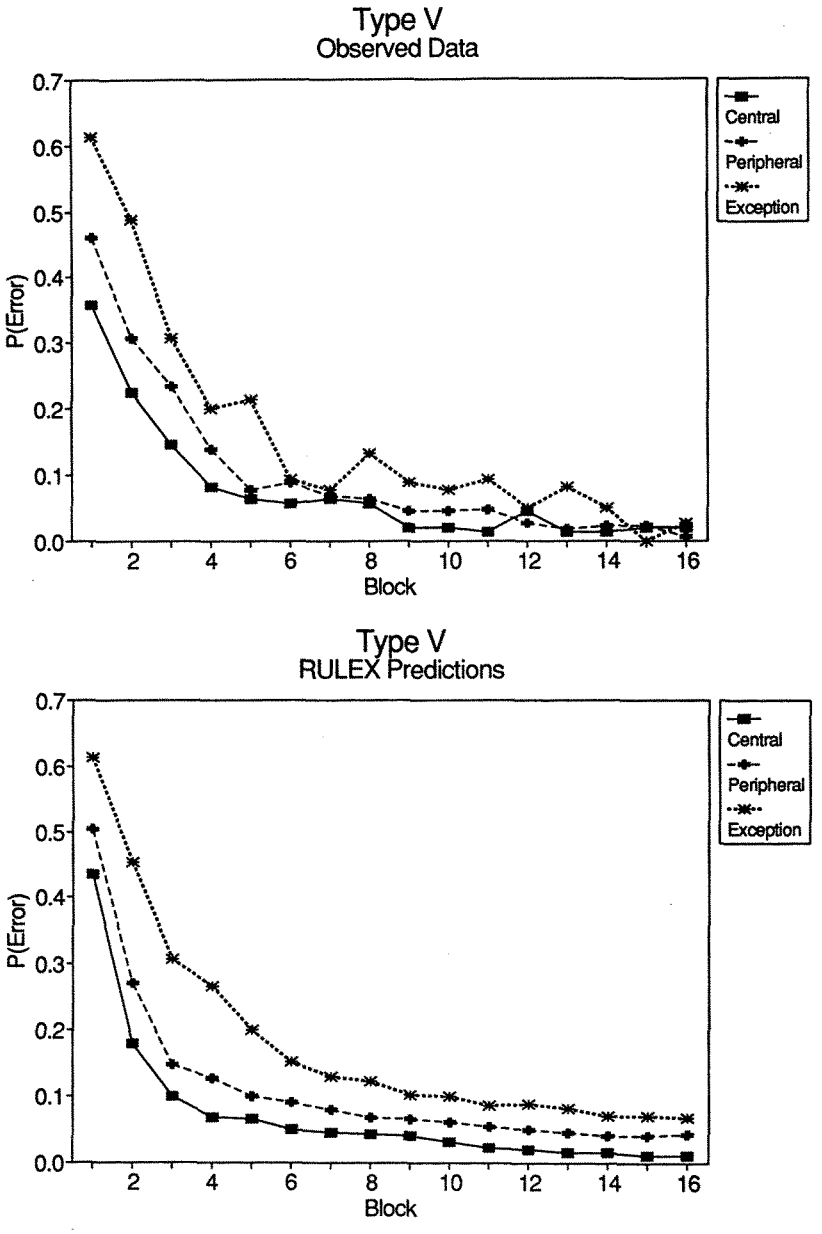


Figure 8. Observed and predicted learning data for the Type V problem. RULEX = rule-plus-exception model of classification learning.

would require an excess number of free parameters or otherwise tailoring the model to that subject's data. This situation is analogous, say, to trying to predict the highly irregular "Brownian motion" that occurs for particles suspended in a liquid. One cannot predict the precise location of any given particle from moment to moment, but elegant mathematical models are available for predicting the overall distribution of locations.[4]

In this section, we report an experiment that was designed to provide more rigorous tests of the ability of RULEX to predict distributions of generalizations. (The experiment discussed in our earlier example involved relatively few subjects, and we needed to cumulate over different types of generalization profiles when evaluating RULEX.) The category structure that we used was the same as in Experiments 2 and 3 of Medin and Schaffer's (1978) study and is shown again in our Table 10. The stimuli were computer-generated drawings of fictitious rocket ships and varied along four binary-valued dimensions (shape of

[4] In arguing for the importance of stochastic learning components, we do not wish to become embroiled in the deeply philosophical debate over whether elementary events in the universe are ultimately deterministic or probabilistic. Given the limited information available at the present level of psychological theorizing, however, positing stochastic components in our learning model is clearly reasonable.

Table 10
Fit of the Context Model and RULEX to the Data From Experiment 1

Stimulus	Observed p	Predicted p: Context	Predicted p: RULEX
	Category A		
A1 1112	.77	.79	.79
A2 1212	.78	.83	.79
A3 1211	.83	.88	.77
A4 1121	.64	.65	.65
A5 2111	.61	.64	.63
	Category B		
B1 1122	.39	.45	.40
B2 2112	.41	.44	.40
B3 2221	.21	.23	.21
B4 2222	.15	.16	.19
	Transfer		
T1 1221	.56	.62	.58
T2 1222	.41	.47	.47
T3 1111	.82	.85	.79
T4 2212	.40	.45	.45
T5 2121	.32	.34	.33
T6 2211	.53	.61	.56
T7 2122	.20	.22	.22

Note. RULEX = rule-plus-exception model of classification learning. Entries are the observed and predicted probabilities with which each stimulus was classified in Category A during the test phase.

the tail, wings, nose, and porthole). Unlike the procedure used in Medin and Schaffer's study, assignment of physical dimensions to the abstract category structure was randomized for each subject so that physical and logical dimensions were not confounded. This procedure allowed us to fit RULEX to the classification data with few free parameters because the weights for representing dimensional salience were not needed. As will be seen, this procedure also allowed us to study whether RULEX could explain certain selective-attention effects that have been reported in the literature.

Experiment 1

Method

Subjects. The subjects were 227 undergraduates from Indiana University, who participated as part of an introductory psychology course requirement. All subjects were tested individually.

Stimuli and apparatus. The stimuli were computer-generated line drawings of fictitious rocket ships varying on four binary-valued dimensions: shape of (a) tail, (b) wings, (c) nose, and (d) porthole. The assignment of physical dimensions and values on dimensions to the abstract category structure (see Table 10) was randomized for each subject. CompuAdd-386 computers were used to generate the stimuli and control the experiment.

Procedure. There were 16 blocks of nine trials. Each of the 9 training stimuli was presented once in each block. The order of training stimuli within each block was randomized. On each trial, a subject judged whether the stimulus was a rocket ship from Planet A or from Planet B, and feedback was then provided. After the training phase, a transfer phase was conducted in which all 16 stimuli were presented. There were 3 blocks of transfer trials, with each stimulus presented once in each block. Subjects judged whether the stimulus belonged to Planet A or Planet B. No feedback was presented during the transfer phase.

Results and Theoretical Analysis

The transfer data are reported in Table 10 in terms of the average probability with which each stimulus was classified into Category A. The overall pattern of data is similar to what Medin and Schaffer (1978) reported in their earlier studies. For example, the prototype of Category A, 1111, was classified in Category A with very high probability, although it was never seen during the learning phase. The Category B prototype, 2222, was also classified with very high accuracy. In addition, the training pattern A2 (1212) was classified in Category A with slightly higher probability than was the training pattern A1 (1112), replicating the effect of specific exemplars on performance.

In an initial theoretical analysis, we fitted Medin and Schaffer's (1978) context model to these classification data. On the basis of previous results, we expected the context model to provide a good fit. One purpose in fitting the context model was to establish a benchmark against which RULEX could be compared. A second purpose was to study the relation between the best fitting parameters in the context model and the rules extracted by RULEX under conditions in which the intrinsic salience of the component dimensions was equated.

According to the context model, the probability that stimulus i (S_i) is classified in Category A is found by summing the similarity of S_i to all exemplars of Category A, and then dividing by the summed similarity of S_i to all exemplars of both Categories A and B. The similarity between S_i and exemplar j (E_j) is given by the multiplicative rule

$$s_{ij} = \Pi s_m^{\delta_m(i,j)},$$

where s_m ($0 \leq s_m \leq 1$) is a free parameter representing the similarity of mismatching values on dimension m and $\delta_m(i, j)$ is an indicator variable set equal to one when S_i and E_j have mismatching values on Dimension m and set equal to zero otherwise. In the present case, the stimuli vary on four dimensions, so the context model has four free parameters, one similarity parameter for each dimension. Under conditions in which the dimensions have equal intrinsic salience, smaller values of similarity in the context model represent greater "attention" devoted to a dimension.

The fit of the context model to the averaged transfer data is shown in Table 10. As expected, the context model provides a good fit to these averaged data, SSD = .030, $RMSD$ = .043, %var = 96.1. The best fitting parameters were s_1 = .267, s_2 = .671, s_3 = .279, and s_4 = .511. Thus, according to the context model, subjects devoted the most attention to Dimensions 1 and 3 and the least attention to Dimension 2. In a previous theoretical analysis, Nosofsky (1984) showed that this pattern of selective attention is in the direction of optimizing subjects' classification performance.

In our next theoretical analysis, we fitted a four-parameter version of RULEX to the classification transfer data. Instead of using point estimates of general storage probability and strict criterion, we used interval estimates. The value of general stor-

age probability varied uniformly from .35 to .65, and the value of the strict criterion for forming a permanent, imperfect, single-dimension rule varied uniformly from .70 to .85. All other parameters were set at their default values, including the dimension-salience weights. The predictions of this four-parameter RULEX model are shown in Table 10. As was the case for the context model, the quantitative fit is very good (*SSD* = .015, *RMSD* = .031, %var = 98.0).

Results of the RULEX simulations indicated that roughly 35% of the rules were based on Dimension 1, 35% on Dimension 3, and 5% on Dimension 4. (Single-dimension rules based on Dimension 2 were extremely rare.) The remaining 25% of the simulations resulted in idiosyncratic rules based entirely on the storage of higher order exceptions.

It is interesting to note the correspondence between the best fitting similarity parameters for the context model and the rules extracted by RULEX. As noted previously here, according to RULEX, the most common single-dimension rules were on Dimensions 1 and 3. According to the context model, these dimensions received the most attention. Conversely, according to RULEX, single-dimension rules were almost never extracted on Dimension 2. According to the context model, this dimension received the least attention. According to both models, then, greater attention is given to the more highly diagnostic dimensions (1 and 3). In RULEX, however, attention corresponds to those dimensions that participate in the simple rules, whereas in the context model attention affects similarities among the exemplars.

The main results of interest in this experiment are shown in Figure 9, which displays the distribution of generalizations observed at the individual subject level. Because there were 7 transfer stimuli and 2 categories, there are $2^7 = 128$ possible patterns of generalization, although only 36 of these patterns were exhibited by at least 2 subjects. For example, the pattern AAABBBB, which was exhibited by 32 subjects, corresponds to those subjects who classified Transfer Stimuli 1 to 3 in Category A and Transfer Stimuli 4 to 7 in Category B. (Because there were three transfer blocks, a subject is said to classify a stimulus into Category A during the transfer phase if he or she classifies it into Category A in at least two of the three blocks.) Two of the most common generalizations were AAABBBB and BBAABAB. The former is consistent with the Dimension 1 rule: 1*** → A, 2*** → B; the latter is consistent with the Dimension 3 rule: **1* → A, **2* → B. A third common generalization was ABABBAB. We discuss possible bases for this generalization later.

Our central goal was to test how well RULEX could account for the observed distribution of generalizations. In addition, we were interested in testing the ability of RULEX to account for certain patterns of response consistency across the three transfer blocks, which we describe shortly. Because we were now fitting these detailed aspects of individual subject behavior, we decided to allow the response-error parameter to vary, in addition to obtaining interval estimates of general storage probability and single-dimension rule criterion. (The response-error parameter turns out to be important for predicting our subsequent response-consistency data but adds very little to the fit of the model to the distribution of generalizations.) With general storage probability varying uniformly from .30 to .65, single-dimension rule criterion varying uniformly from .65 to .85, and a response error of .07, the distribution of generalizations predicted by RULEX is shown in Figure 10. This version of RULEX accounts for 85.7% of the variance in the distribution of generalizations. Moreover, with these parameters, RULEX accounts for an impressive 99.0% of the variance in the complete set of averaged transfer data (see Table 10). Thus, in addition to accounting extremely accurately for the averaged transfer data, RULEX does at least a fair job of simultaneously characterizing the patterns of variability existing at the individual subject level. Among RULEX's main shortcomings, however, is that it underpredicts the frequency of generalization ABABBAB.

As a source of comparison, we also fitted the context model to the distribution of individual subject generalizations. The best fit of the context model is illustrated in Figure 11. As is evident from inspection, the context model does far worse than RULEX at predicting the distribution of generalizations (%var = 35.9). Among its main shortcomings is that it underpredicts the frequency of the single-dimension generalizations (AAABBBB and BBAABAB).[5]

Although its fit to the overall distribution of generalizations is relatively poor, it is intriguing that, unlike RULEX, the context model correctly predicts that generalization ABABBAB should occur with high frequency. Indeed, if one assumes the parameters that maximize the fit of the context model to the averaged classification data, then the pattern ABABBAB is predicted to be the highest frequency generalization. One interpretation for this pattern of generalization, therefore, is that it reflects subjects who were using an exemplar storage strategy to learn the classification problem and basing classification decisions on the similarity of objects to these exemplars. Although speculative, this analysis provides preliminary support for the idea that a mixture of classification strategies may have occurred in this experiment, with some subjects using something akin to a rule-plus-exception strategy and others using an exemplar-storage strategy.

In another theoretical comparison, we tested RULEX and the context model on their ability to predict the degree of response consistency in individual subjects' classification decisions. As a measure of response consistency, we computed the "hamming distance" between generalization profiles exhibited by a subject in Transfer Blocks 1 and 2. The hamming distance is the number of transfer stimuli for which the subject gave inconsistent responses across blocks. For example, if the subject exhibited generalization profile AAAAAAA during Block 1 and generalization profile BBAAAAA during Block 2, then the hamming distance for that subject was 2.

---

[5] Because it uses a probabilistic response rule, the predicted distribution of generalizations from the context model was computed from the average probability vector produced by the model. For example, let $p1$ denote the predicted probability that Transfer Stimulus 1 is classified in Category A. Then the proportion $p1$ of the generalization profiles had a value of 1 in their first position, and the proportion $1\text{-}p1$ of the generalization profiles had a value of 2 in their first position. An additional method for producing variability at the individual subject level is to develop stochastic versions of the context model analogous to the way that RULEX is formalized, but no theories along these lines have yet been proposed.

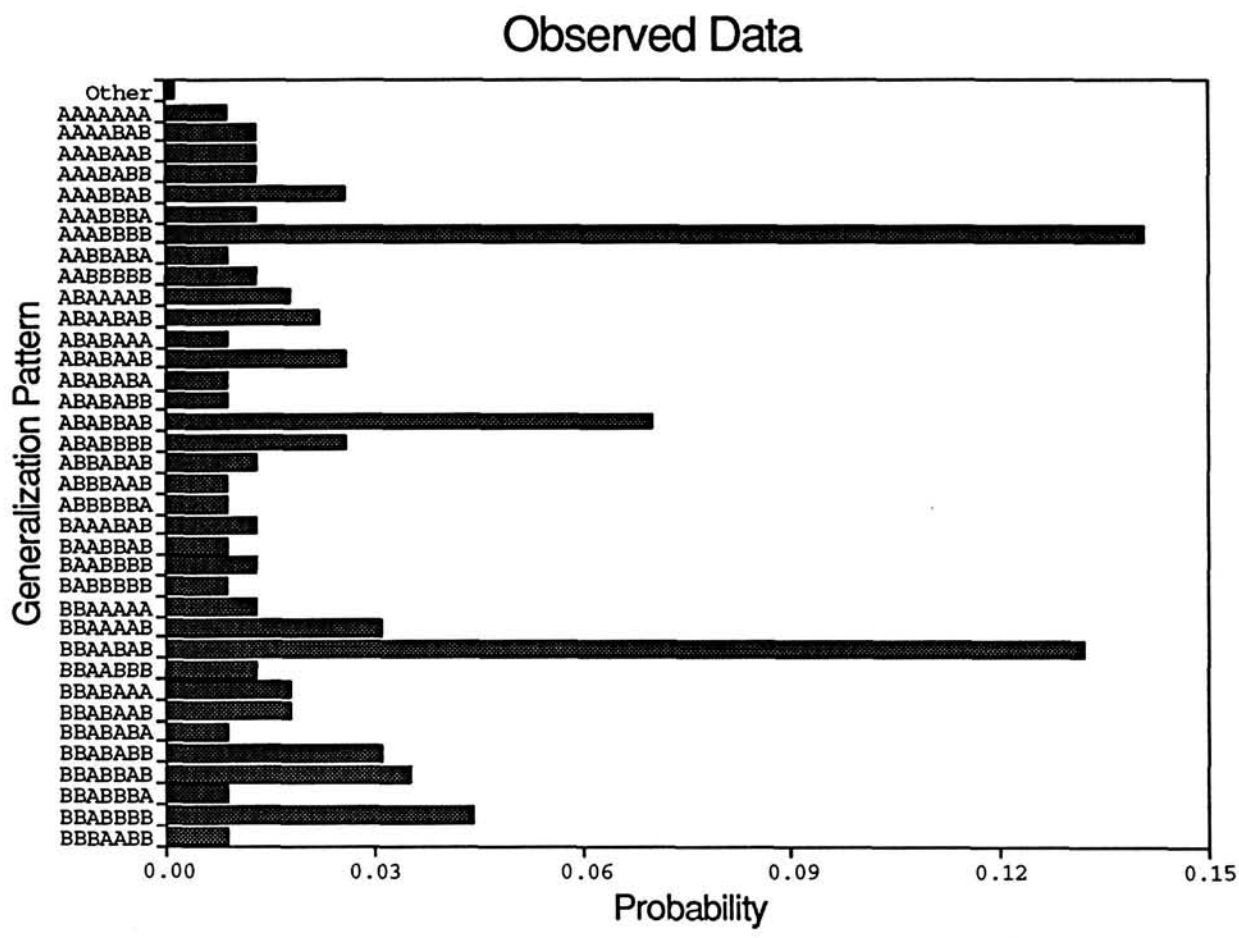


Figure 9. Observed distribution of generalizations for Experiment 1.

In Figure 12 (Panel A), we plot the observed distribution of hamming distances between the Block 1 and Block 2 generalization profiles for all 227 subjects. (The distributions of hamming distances were similar between Blocks 2 to 3 and 1 to 3.) The most common hamming distances were 0 and 1, and the frequency of hamming distances declined steadily thereafter. Also shown in Figure 12 are the predicted distributions of hamming distances from RULEX (Panel B) and the context model (Panel C). The parameters that were used in generating these predictions were held fixed at those values that provided a best fit to the distribution of generalizations. It is evident from inspection that, whereas RULEX does a fair job of characterizing the degree of response consistency across blocks of transfer, the predictions of the context model are quite poor. It is important to acknowledge, however, that the response-error parameter in RULEX is critical in allowing that model to describe the degree of response consistency. If the response-error parameter is set at zero, RULEX's fit to the distribution of generalizations (see Figure 9) is essentially as good as before, but it predicts too much response consistency across blocks of transfer. Another interpretation for the distribution of hamming distances, consistent with our interpretation of the distribution of generalizations, is that it may reflect a mixture of classification strategies. If the predictions of RULEX, with, say, a response error of .02, are mixed with the predictions of the context model, a good fit to the distribution of hamming distances can also be achieved.

The differing predictions of response consistency from RULEX and the context model stem from the different response rules found in these two models. Whereas the response rule in RULEX is nearly deterministic, the context model uses a probabilistic response rule (see footnote 5). Apparently, individual subjects in the present experiment responded in a more consistent manner than is predicted by the probabilistic response rule in the context model. This finding agrees with an earlier one reported by Ashby and Gott (1988) in a much different experimental paradigm. As noted by Nosofsky (1991b), it is possible to modify the context model by using a deterministic response rule and introducing noise in other locations. It may also be possible to develop stochastic learning versions of the context model, but such undertakings are beyond the scope of the present article. Our limited conclusion is simply to note that the present model, RULEX, appears to do a fair job of characterizing patterns of generalizations and response consistency existing at the individual subject level, and the results pose an interesting challenge for some extant models.

In summary, a key contribution of these analyses is the demonstration that, although classification models may account extremely accurately for averaged data, vast individual differences

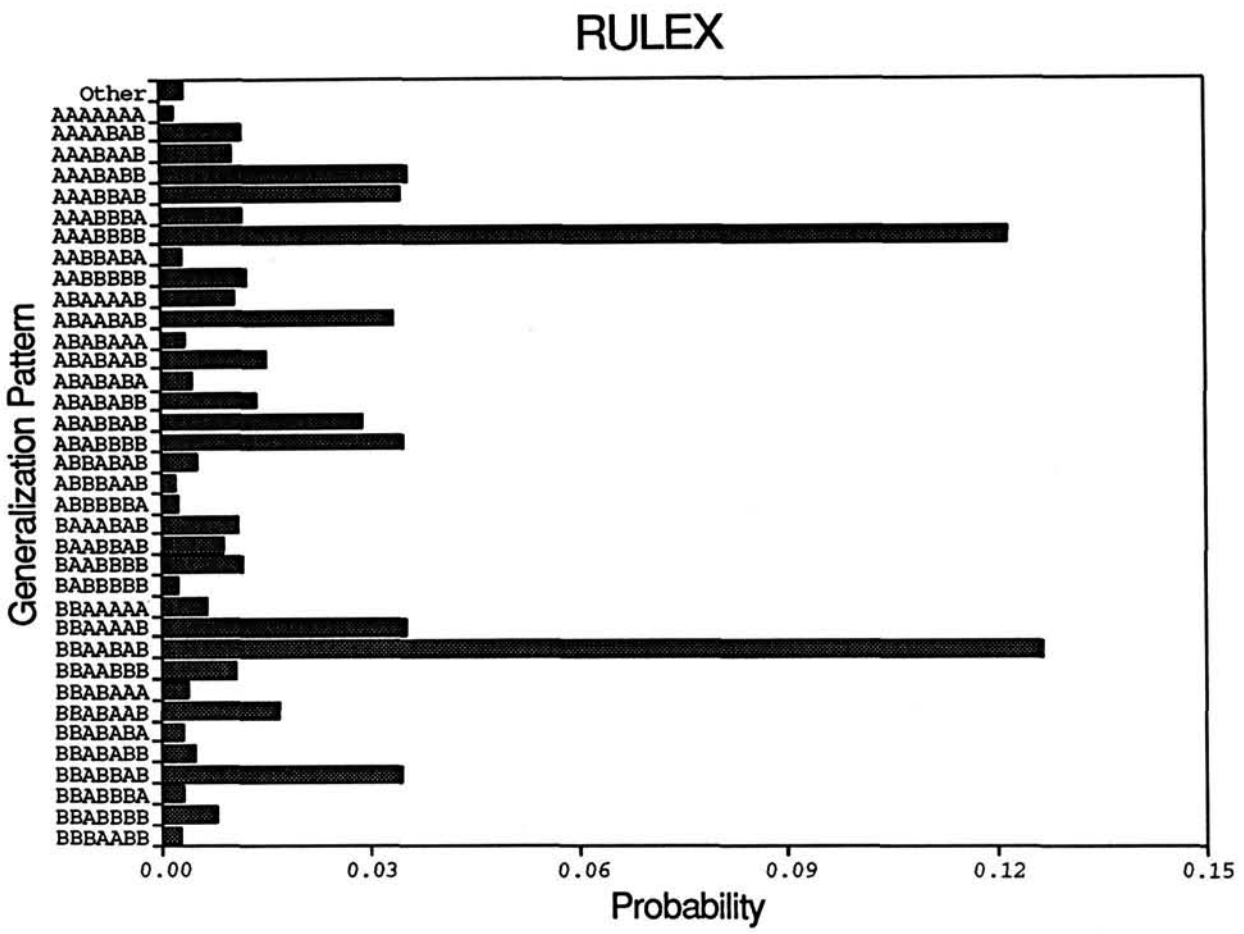


Figure 10. Predicted distribution of generalizations from the rule-plus-exception (RULEX) model.

may lurk beneath the surface. Of course, the basic idea that aggregate data may not reflect patterns observed at the individual subject level is extremely well known. The present innovation, however, is to characterize the range of individual differences in categorization in terms of distributions of generalizations and measures of response consistency and to test models on their ability to account for these distributions of individual subject behaviors. Although there is room for improvement, we believe that the present model makes some significant headway toward simultaneously characterizing the aggegrate classification data and the individual subject behaviors from which the aggregate data are derived.

General Discussion

In this article, we investigated the idea that, in situations involving highly separable-dimension stimuli, much of category learning may involve the extraction of simple logical rules, with occasional exceptions to those rules also being stored. This idea was formalized within the framework of a computer simulation model of rule-plus-exception learning (RULEX). RULEX was able to account for a variety of classic phenomena reported in the categorization literature. It accounted simultaneously for the prototype and specific exemplar effects reported in Medin and Schaffer's (1978) studies; for people's sensitivity to both correlated and individually diagnostic dimensions, and how this sensitivity evolves as a function of category learning; for the relative ease of learning nonlinearly separable versus linearly separable categories; for the speed of learning the six problem types in the classic studies of Shepard et al. (1961); and for selective-attention phenomena. We also demonstrated that RULEX may contribute to a fuller understanding of the vast individual differences in patterns of generalization that exist at the individual subject level. Beyond accounting for these effects at a qualitative level, we demonstrated that RULEX is often able to achieve accurate quantitative accounts of these phenomena. Thus, the idea that people learn categories by forming simple rules and storing occasional exceptions appears to be viable and deserving of continued investigation. Apparently, simple hypothesis-testing and rule-extraction processes may have much more generality and explanatory power for classification learning than previously imagined.

Limitations and Extensions

RULEX was intended to be a fairly simple representative of a more general rule-plus-exception model. We believe that RULEX will need to be extended in various ways to provide more complete accounts of category learning and representation. Of course, this preliminary version of the model is appli-

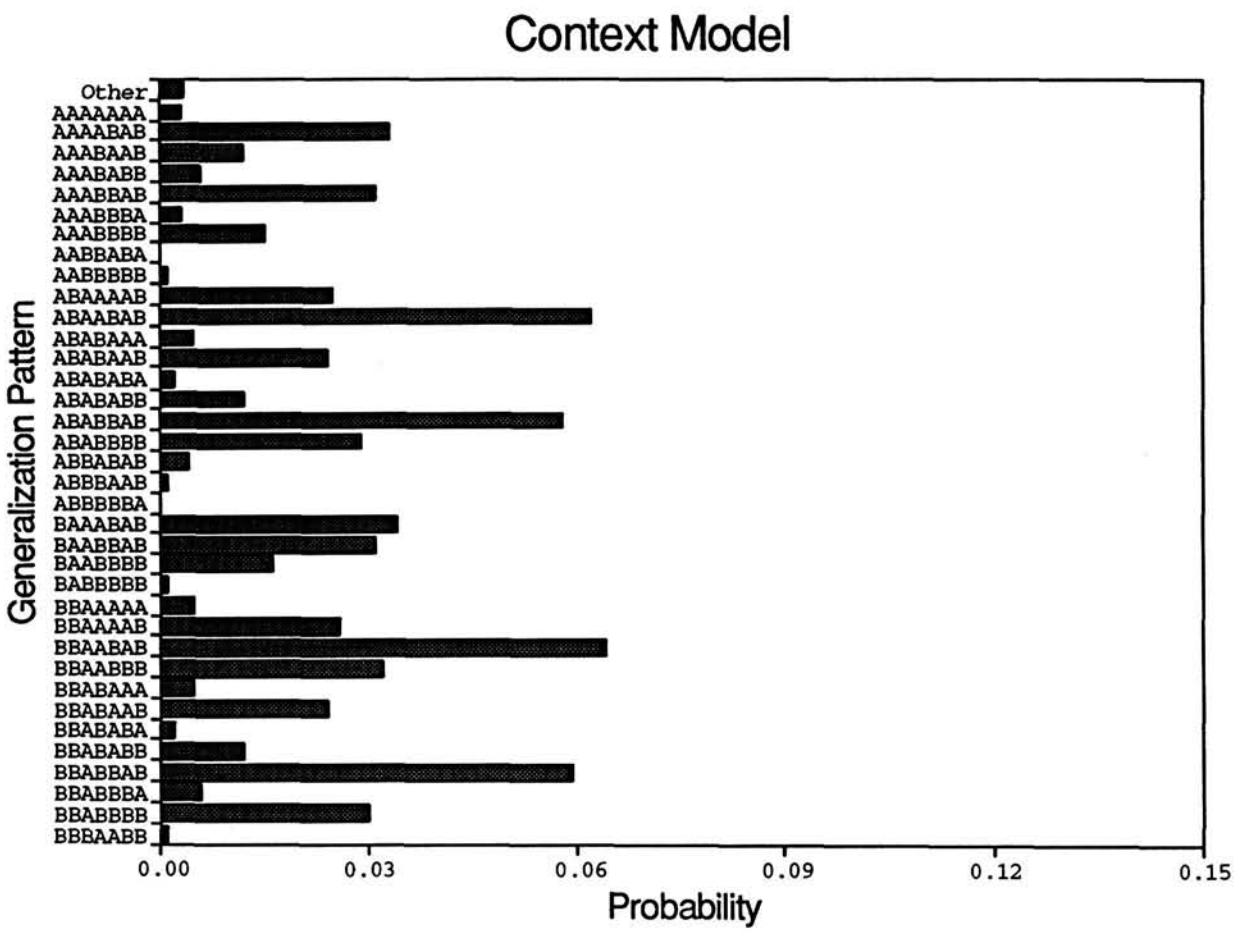


Figure 11. Predicted distribution of generalizations from the context model.

cable only in situations involving binary-valued dimensions, two categories, and deterministic assignments of exemplars to categories, so it clearly needs to be generalized along these lines.

We also believe that the current version of RULEX is too brittle, and more flexibility needs to be built into the system. For example, rules are tested and formed in a rigid, sequential manner. Perfect single-dimension rules are tested first, followed by imperfect single-dimension rules, and finally conjunctive rules. It seems likely that stochastic mechanisms need to be added to the system to allow for more variable sequencing and for the extraction of highly accurate, albeit imperfect rules at the outset. In addition, in the current version of RULEX, once an imperfect rule has been discarded, it is no longer available for later use. It seems likely that a subject might return to such a rule if he or she eventually discovered that it was the best among several other alternatives. Allowing the model to go back to testing old rules or to storing multiple alternative rules in parallel also seems important for handling experimental paradigms in which exemplar-category assignments are shifted midstream. Another aspect of RULEX that is probably too inflexible is its strict all-or-none criterion for applying rules and exceptions. For example, in the current version of RULEX, if an object is highly similar to an exception but not identical to it, the exception is not applied. An intriguing possibility for extending RULEX involves the use of adaptive network models that dynamically select alternative rules and exceptions to test and apply (Busemeyer & Myung, 1992; Choi, McDaniel, & Busemeyer, 1993; Kruschke, 1992).

Despite these limitations, we find it instructive that even a simple representative of a rule-plus-exception model can account for the wide variety of categorization phenomena that we investigated in this article.

Old–New Recognition and Exemplar Memories

Another question concerns how RULEX would be applied to handle old–new recognition judgments. Often, after the completion of category learning, subjects show an ability to discriminate between old and new exemplars. Furthermore, good accounts of the patterns of old–new recognition data are often provided by summed-similarity exemplar models (e.g., Estes, in press; Hintzman, 1988; Medin, 1986; Nosofsky, 1988, 1991b). According to these models, recognition judgments are based on overall familiarity, in which familiarity is computed by summing the similarity of an item to all category exemplars stored in memory.

If all that is stored in memory is a rule and a few exceptions,

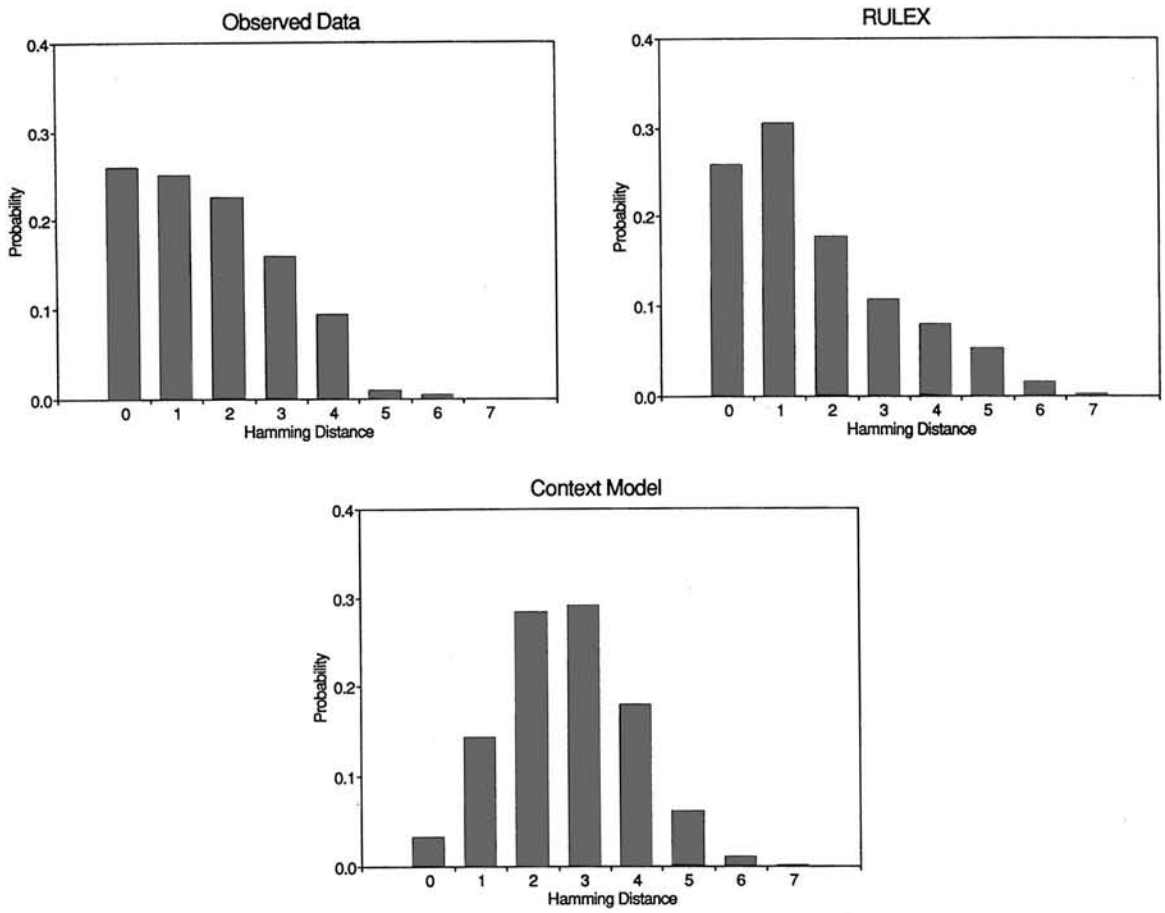


Figure 12. Panel A: Observed distribution of hamming distances for Experiment 1. Panel B: Predicted distribution of hamming distances from the rule-plus-exception (RULEX) model. Panel C: Predicted distribution of hamming distances from the context model.

how does this ability to discriminate between old and new exemplars emerge, and why do summed-similarity exemplar models yield good fits? One answer to this question is to admit that, in addition to forming rules and exceptions, some subjects may also have some residual memory for the exemplars that were presented during training. Recognition judgments could then be based on these exemplar memories, even if the dominant strategy for solving classification problems is to form rules. If this suggestion is correct, then an interesting issue is whether RULEX could be combined with exemplar models to yield still more accurate predictions of recognition judgments.

Our main idea is that the exceptions that subjects store when solving a classification problem may be the most memorable old exemplars (or parts of exemplars). We have preliminary data that strongly support this idea. Following a procedure used by Medin and Smith (1981), we gave subjects explicit instructions to use a particular single-dimension-plus-exception strategy to solve a classification problem. We collected old–new recognition judgments at the end of classification learning. The items with the highest recognition probabilities were, by far, those items that were the exceptions to the single-dimension rule. Other items with high recognition probabilities were those that were highly similar to the exceptions. An excellent account of the old–new recognition data was provided by a summed-similarity model that combined weak memories for the old exemplars with strong memories for the exceptions. We are currently pursuing this idea of combining RULEX with exemplar models to predict old–new recognition judgments in free-strategy situations.

More generally, a critical issue for future research is to understand the combined role of rules and exemplars in categorization and memory and the experimental conditions that promote the use of these alternative strategies (e.g., Brooks, 1987; Nosofsky et al., 1989; Ward & Scott, 1987; Wattenmaker, 1993). Even when fairly simple logical rules are available, there is evidence that exemplars may play a critical role in classification (Allen & Brooks, 1991; Nosofsky et al., 1989), typicality judgments (Nosofsky, 1991c), and implicit memory phenomena (Whittlesea, 1987). This evidence for the coexistence of rules and exemplars converges with the finding in the present study that, although RULEX accounted for a substantial proportion of the variance in the distribution of individual subject generalizations, at least some of the remaining patterns of generalization could be interpreted as exemplar based.

Relations to Other Models

In this section, we briefly discuss the relation between RULEX and some other models of classification.

Context model. Although our intent was to establish a contrast between RULEX and the exemplar-similarity context model, it is important to acknowledge that these models share some important commonalities (cf. Medin, 1986). Both models are relational coding models in the sense that they will display sensitivity to co-occurrence information in the features that compose the exemplars. RULEX displays this sensitivity because the exceptions that it stores correspond to bundles of co-occurring features, and the context model displays this sensitivity because there is a memory representation for each of the exemplars. RULEX and the context model also have in common the assumption that certain dimensions are more salient in the category representation than are others. In RULEX, those dimensions that have intrinsic salience or that are highly diagnostic are more likely to participate in the single-dimension or conjunctive rules that are formed. In the context model, greater attention is given to these dimensions, which affects similarity relations among the stored exemplars. Indeed, because of these commonalities, Medin suggested that exemplar models that include assumptions about selective attention (such as the context model) may often provide good fits to data associated with the use of rule-based strategies.

ACT generalization model. Anderson et al. (1979) and Elio and Anderson (1981) presented a model of classification learning that operated according to the principles of production-rule formation in the ACT system (Anderson, 1976). This model has some similarities to RULEX in that both general and specific rules may coexist by the time a subject has solved a classification problem. In addition, the rule systems that are formed may vary across different subjects, so the model's predictions of aggregate data will not match those for any given individual subject. There are several major conceptual differences between RULEX and the ACT generalization model, however. In RULEX, we assume that subjects first develop very simple rules by means of hypothesis testing and then gradually supplement these simple rules with exceptions. The opposite learning process tends to occur in Anderson et al.'s model. Subjects start by storing complete exemplars in memory and subsequently form generalizations that reflect commonalities among exemplars from the same category. Thus, this system presumes a good deal of memory for past examples during the learning process. Indeed, in their tests of the ACT generalization model, Elio and Anderson reported that their subjects were "encouraged not to formulate and test hypotheses during learning, but to concentrate on memorizing each description . . . we stressed item memorization rather than active hypothesis testing" (pp. 403–404). In addition to these fundamental differences in the learning process, RULEX and the ACT generalization model differ in their view of the final category representation and decision process. In RULEX, we suppose that each subject has stored at most a couple of simple rules together with a few exceptions, and these rules are applied in an essentially deterministic manner. By contrast, in the ACT generalization model, numerous rules exist in parallel, and the particular rule that is selected for application on a given trial is chosen probabilistically.

Configural cue model. Gluck and Bower's (1988) configural cue model is similar to RULEX in that associations are learned between both single features and categories and between configurations of features and categories. By the time category learning is completed, the representation in the configural cue model might be similar to what occurs in RULEX. For example, a strong association might exist between a single feature and the alternative categories, but this single-feature association might be supplemented by associations between some highly specific configurations of features (exceptions) and the categories. RULEX and the configural cue model differ, however, in their conceptualization of the learning process. In RULEX, relations between features and categories are learned in an all-or-none manner by a hypothesis-testing process, and only a single association is learned at a time. By contrast, in the configural cue model, associations are learned by gradually incrementing and decrementing a set of connection weights, and all associations are learned in a simultaneous and interactive manner. Furthermore, whereas the configural cue model presumes that a massive number of associations are learned between all possible feature combinations and the categories, RULEX supposes that only a few such associations are learned by any given subject. Finally, whereas the learning process in RULEX is highly stochastic and leads the model to predict vast individual differences at the individual subject level, current versions of the configural cue model use a deterministic learning process.

General recognition theory (GRT). Ashby and Townsend's (1986) GRT is a multidimensional generalization of signal-detection theory. According to the theory, the observer establishes decision boundaries in a psychological space that partition the space into response regions. Any internal representation that falls in Region A would result in a Category A response. As discussed by Ashby (1992), the GRT provides a very general framework in which numerous different models of classification can be expressed, depending on the types of decision boundaries that are assumed. In some applications involving stimuli varying along two continuous dimensions, Ashby and Lee (1991) and Maddox and Ashby (1993) successfully fitted GRT models that assumed general linear boundaries, quadratic boundaries, or exemplar-based likelihood boundaries, although no process model was tested for the learning of these boundaries. In the language of Ashby and Gott (1988) and Ashby (1992), the rules and exceptions developed by RULEX could be characterized as complex sets of independent-decisions boundaries. RULEX provides an explicit process model for how such decision boundaries are learned over time (in stimulus domains with binary-valued dimensions) and provides a theory of the range of differences in types of decision boundaries observed across individual subjects.

Two-stage model of category construction. Whereas the focus of the present research was on classification learning, an important, closely related cognitive process is category construction. In a category-construction task, subjects are given a number of objects and are asked to cluster them into whatever groups seem most natural. Given the enormous evidence that natural categories are defined by similarity relations, or family resemblance, it is reasonable to expect that people would cluster objects in a free-sorting task on the basis of family resemblance as well. However, evidence suggests that people most often sort

objects on the basis of a single salient dimension (Ahn & Medin, 1992; Medin, Wattenmaker, & Hampson, 1987; Wattenmaker, 1992). Furthermore, Ahn and Medin found that occasional cases of apparent family resemblance sorting could be accounted for by a two-stage model of category construction. In the first stage, objects are clustered on the basis of a single dimension. In the second stage, the exceptions, which cannot be clustered on the basis of this dimension, are placed into the group to which they are most similar. Although the task goals are quite different and learning processes based on hypothesis testing are not involved, this two-stage model of category construction bears a striking resemblance to RULEX. Ahn and Medin's (1992) work can thus be viewed as providing still further evidence about the potential generality of rule-plus-exception processes in categorization.

Rule-plus-exception models. RULEX is closely related to a variety of other rule-plus-exception models, mainly from the artificial intelligence and machine-learning literatures (e.g., Fisher, 1987; Hunt et al., 1966; Medin, Wattenmaker, & Michalski, 1987; Michalski, 1983; Quinlan, 1986; Schlimmer, 1987). Like RULEX, such algorithms are designed to formulate a fairly simple system of rules for correctly classifying objects into alternative categories. One of the main features that distinguishes RULEX from many of these other models, however, is that RULEX is intended to be a psychologically plausible learning model. By contrast, many of these other models are more concerned with designing algorithms that will, in some sense, construct optimal systems of rules. Other such models either are not concerned with learning per se or place large memory demands on the concept-learning system. For example, the INDUCE and PATCH models of Michalski (1983) and Medin, Wattenmaker, and Michalski (1987), which have had some interesting psychological applications, construct systems of rules for classifying examples in situations in which multiple training exemplars are simultaneously present. The same is true of Quinlan's (1986) ID3 model, which must examine and reexamine all previously presented exemplars at many stages of learning. Even the classic concept-learning systems developed by Hunt et al. (1966) presumed that there was memory either for all previously presented exemplars or for a substantial subset of them. Furthermore, after any classification error, all rules were discarded, and an entirely new system of rules was constructed in toto to classify correctly all exemplars residing in memory. By contrast, in RULEX, the rule-induction process takes place on a trial-by-trial basis by means of hypothesis testing, and there is only limited memory for any exemplars presented on previous trials. There are some examples of trial-by-trial rule-learning algorithms from artificial intelligence, but, rather than being based on explicit hypothesis-testing procedures involving limited memory, such algorithms presume much richer sources of statistical information involving feature-category correlations, measures of category utility, and so forth (e.g., Fisher, 1987; Schlimmer, 1987).

Regardless of how one views such differences, however, the most important novel contribution of the present work is that we have demonstrated the ability of a rule-plus-exception model to account for a wide variety of categorization phenomena of major psychological interest and have provided quantitative tests of the ability of such a model to predict psychological data.

Directions for Future Research

The main theme of our investigation was to demonstrate the ability of RULEX to account for well-known results in the classification literature. These demonstrations are important because the learning process and category representation in RULEX differ substantially from other current models in the field, most of which have extraordinarily high information-processing demands. The next logical step is to begin to develop contrasts that will allow the predictions of RULEX to be tested against those of competing models. We suspect that the best chance of developing such contrasts will involve detailed analyses of learning and transfer data at the individual subject level. For example, early research on concept formation provided evidence that individual subjects learn simple rules in an all-or-none manner (Trabasso & Bower, 1968). Such all-or-none learning phenomena are consistent with the hypothesis-testing process that is assumed in RULEX but may be much more difficult for incremental learning models such as ALCOVE, the configural cue model, and the rational model to explain. Likewise, instead of comparing models on their ability to predict averaged classification transfer data, it may prove more diagnostic to compare them on their ability to predict patterns of individual subject variability, such as the distributions of generalizations that we have considered herein. Before rigorous comparisons can be achieved, however, it will be important to extend essentially all of these by models by incorporating stochastic learning components. Such components are probably necessary to handle the extensive variability in strategies that appears to exist at the individual subject level. This need to rely on more and more fine-grained levels of analysis attests to the theoretical progress being made in the field of category learning and representation.

References

Ahn, W. K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120,* 3–19.

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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* (pp. 277–318). San Diego, CA: Academic Press.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

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

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categori-

zation from identification. *Journal of Experimental Psychology: General, 120,* 150–172.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: The ecological and intellectual factors in categorization* (pp. 141–147). Cambridge, England: Cambridge University Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121,* 177–194.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Evaluation of exemplar-based and network models of conceptual rule learning. *Memory & Cognition, 21,* 413–423.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 397–418.

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

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

Estes, W. K. (in press). *Classification and cognition.* Oxford, England: Oxford University Press.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2,* 139–172.

Fowler Williams, G. (1971). A model of memory in concept learning. *Cognitive Psychology, 2,* 158–181.

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

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

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

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction.* San Diego, CA: Academic Press.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science, 5,* 3–36.

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology, 71,* 331–338.

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.

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.

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

McKinley, S. C., & Nosofsky, R. M. (1993). *Attention learning in models of classification.* Manuscript in preparation.

Medin, D. L. (1986). Commentary on "memory storage and retrieval processes in category learning." *Journal of Experimental Psychology: General, 115,* 373–381.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

Medin, D. L., & Florian, J. E. (1992). Abstraction and selective coding in exemplar-based models of categorization. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 207–234). Hillsdale, NJ: Erlbaum.

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 and Memory, 7,* 355–368.

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

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

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.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence, 20,* 111–161.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology, 64,* 640–645.

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

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

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

Nosofsky, R. M. (1991a). Relation between the rational model and the context model of categorization. *Psychological Science, 2,* 416–421.

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

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

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, pp. 149–168). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. T. (in press). Comparing models of rule-based classification

learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition.*

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.* Hillsdale, NJ: Erlbaum.

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

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1,* 81–106.

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

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

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

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

Schlimmer, J. C. (1987). Incremental adjustment of representations for learning. *Machine Learning, 2,* 79–90.

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. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

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

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

Wattenmaker, W. D. (1992). Relational properties and memory-based category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 203–222.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 3–17.

Appendix

We now provide some additional details concerning the RULEX simulation.

During the search for imperfect single-dimension rules, we assume that the subject establishes a counter on each tested dimension and uses a majority rule to make classification decisions. For example, suppose that the subject samples Dimension 1 and finds that Value 1 on this dimension signals Category A. Then the subject would increment the rule counters 1*** → A and 2*** → B, and these would be the tentative rules. If the next two stimuli had a value of 1 on Dimension 1 but signaled Category B, however, then the alternative counters would be incremented, and the tentative rules would be switched to 1*** → B and 2*** → A. The storage-probability parameter is also involved in the search for imperfect single-dimension rules. On each trial, the probability that the rule counters are actually incremented is given by the storage-probability parameter. This assumption is intended to approximate the idea that people may not have perfect memories for the single-dimension information that is experienced during training.

The search for conjunctive rules operates in a similar manner. Suppose that the subject is sampling Dimensions 1 and 2 and finds that the combination 11** signals Category A. Then the tentative rule would be 11** → A. In addition, the subject would also form the tentative rules 12** → B and 21** → B. The reason for automatically forming these latter tentative rules is that single-dimension rules have already failed. (For example, if 11** → A *and* 12** → A, then it is logical that 1*** → A.) Thus, the rule counters 11** → A, 12** → B, and 21** → B would all be incremented by 1. There is a question, however, concerning the status of the pattern 22**, which could signal either category. In the current version of the simulation, RULEX does not increment either counter involving this pattern and waits until it receives explicit information concerning the category membership of this pattern. As was the case for the single-dimension counters, RULEX uses a majority rule to make its classification decisions. In addition, the probability that the conjunctive rule counters are actually incremented at each step is given by the storage-probability parameter squared. This assumption is intended to approximate the idea that it is probably more difficult to remember conjunctive rule information than single-dimension information.

As discussed in the text, the probability of successfully remembering exceptions is influenced by two parameters: storage probability and capacity limit. Suppose that a subject samples an exception with n dimensions and there are currently m exceptions already stored in memory. Then the probability that the new exception is successfully stored is given by $\text{pstor}^n \cdot \text{capac}^m$. In general, then, it is more difficult to store exceptions composed of a large number of dimensions and more difficult to store new exceptions when there are already a large number of old exceptions in memory. Again, these assumptions approximate the idea that our memory systems are limited in capacity.

Received December 18, 1992
Revision received July 14, 1993
Accepted August 9, 1993 ■