Journal of Experimental Psychology: General
1998, Vol. 127, No. 2, 107–140

Copyright 1998 by the American Psychological Association, Inc.
0096-3445/98/$3.00

Rules and Exemplars in Category Learning

Michael A. Erickson and John K. Kruschke
Indiana University Bloomington

Psychological theories of categorization generally focus on either rule- or exemplar-based explanations. We present 2 experiments that show evidence of both rule induction and exemplar encoding as well as a connectionist model, ATRIUM, that specifies a mechanism for combining rule- and exemplar-based representation. In 2 experiments participants learned to classify items, most of which followed a simple rule, although there were a few frequently occurring exceptions. Experiment 1 examined how people extrapolate beyond the range of training. Experiment 2 examined the effect of instance frequency on generalization. Categorization behavior was well described by the model, in which exemplar representation is used for both rule and exception processing. A key element in correctly modeling these results was capturing the interaction between the rule- and exemplar-based representations by using shifts of attention between rules and exemplars.

Many formal and folk psychological theories conceive of the mind as being composed of quasi-independent modules. From Freud to Fodor, the mind has been decomposed into constituent parts. Recently, a number of researchers have proposed modular theories of cognitive phenomena such as categorization (Ashby et al., in press; Shanks & St. John, 1994), reasoning (Sloman, 1996), automaticity (Logan, 1988), language (Pinker, 1991), and learning and memory (Squire, 1992). In general, these theories are characterized by descriptions of each module and how each serves in those tasks for which it is best suited. However, these theories often do not emphasize how modules interact in producing responses and in learning.

BACKGROUND

In this article, we develop a modular theory of categorization that follows from two distinct accounts of this behavior. The first account is that of rule-based theories of categorization. These theories emerge from a philosophical tradition in which concepts and categorization are described in terms of definitional rules. For example, if a living thing has a wide, flat tail and constructs dams by cutting down trees with its teeth, then it is a beaver. Wittgenstein (1953) noted that some concepts could not be accounted for by this criterial view of categorization. In particular, he offered the concept of "game" as an instance that could not be adequately described in terms of necessary and sufficient conditions. He proposed instead that concepts are based on family resemblances, which emphasize the importance of *similarity* between the elements of a category rather than necessary and sufficient conditions. Similarity forms the basis for the second account of categorization we consider.

Psychological theories of categorization generally followed from one of these two different philosophical accounts, emphasizing either strict rules or similarity to category instances. Both types of theories may be described within a mutual framework. Given stimuli that can be represented as points in a multidimensional, psychological space, a rule can be represented as a division of that space into appropriate category regions. In a rule-based theory of categorization, a classification is made by considering where a percept falls in the multidimensional space and responding based on the surrounding category region. In exemplar-based theories of categorization, stimuli are also represented as points in a multidimensional space. In these models, however, no boundaries are formed. Instead, the similarity between a given percept and previously stored exemplars is computed. Similarity is represented in these models as a monotonically decreasing function of distance in the psychological space. The probability that the percept is classified as belonging to a certain category is a function of the percept's similarity to exemplars of that category relative to the percept's similarity to exemplars of other categories.

In this article, we generally constrain the use of the term *rule* to refer to a *dimensional boundary*, a boundary that is orthogonal to a psychological dimension (e.g., "Large items are in Category A: small ones are in Category B"; the rule forms a boundary orthogonal to the dimension of size). The types of dimensional rules derived for a set of stimuli, however, will depend on how participants represent its dimensional structure. For example, Krantz and Tversky (1975) showed that rectangles may be represented by

Michael A. Erickson and John K. Kruschke, Department of Psychology, Indiana University Bloomington.

This work was supported by Indiana University Cognitive Science Program Fellowships, by National Institute of Mental Health (NIMH) Research Training Grant PHS-T32-MH19879-03 to and in part by NIMH FIRST Award 1-R29-MH51572-01.

This research was reported as a poster at the 1996 Cognitive Science Society Conference in San Diego, CA. We thank Leola Alfonso-Reese, Nathaniel Blair, Michael Fragassi, Mark Johansen, Robert Nosofsky, and Teresa Treat for helpful criticisms and suggestions.

Correspondence concerning this article should be addressed to Michael A. Erickson, who is now at the Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 or to John K. Kruschke, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent to erickson@cmu.edu or to kruschke@indiana.edu.

participants either along the dimensions of area and shape or along the dimensions of height and width. In this article, we focus on these simple, dimensional rules. We will not generally consider conjunctive or disjunctive combinations of rules because they are beyond the scope of our experiments.

Empirical Evidence for Rule- and Exemplar-Based Theories

Both rule- and exemplar-based theories of categorization have accumulated a wide range of empirical support. One example of evidence supporting rule-based theories was provided by Rips (1989). He gave participants a description of an item, such as "a circular object with a 3-in. diameter," and asked them one of two questions: whether the item was more *similar* to a pizza or to a quarter or whether it was more *likely to be* a pizza or a quarter. In the first condition, they responded that the object was more similar to a quarter, and in the second, they responded that it was more likely to be a pizza. Rips interpreted these results to mean that in the second task, a rule was overriding participants' similarity judgments.

A number of researchers noted that categories typically have "graded" structures (Rips, Schoben, & Smith, 1973; Rosch & Lloyd, 1978; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976). This means that not all members of a category have the same degree of membership. Converging measures identify these differences. For example, better members of a category are identified more quickly and with greater accuracy. Moreover, participants in the experiment can often state explicitly which instances are more or less typical of a given category. These facts, by themselves, do not distinguish between rule- and exemplar-based theories. To accomplish this, further analysis of the nature of the category structures is necessary. Under rule-based theories, gradation must be computable using two pieces of information: a percept and a category boundary. In contrast, a percept and all previous instances are available for use in exemplar-based theories.

Effects of the distance of a percept from a category boundary are seen in experiments in which a single rule can be used to distinguish the members of each category. Imagine that participants are instructed to group circles larger than 3 cm into one category and circles smaller than 3 cm into another; they will be more accurate and faster when classifying 1-cm and 6-cm circles than when classifying 2.9-cm and 3.1-cm circles. Hence, category membership is improved as the distance from the category boundary increases. Such results can be explained by both rule-based (see, e.g., Ashby & Lee, 1991, 1992, 1993; Ashby & Maddox, 1992, 1993) and exemplar-based (Nosofsky, 1986, 1987, 1988b, 1989) theories. (For a recent study demonstrating the inadequacy of rule-only accounts of categorization of one-dimensional stimuli, see Kalish & Kruschke, 1997.)

Brooks and colleagues (Allen & Brooks, 1991; Regehr & Brooks, 1993), however, used more complex category designs to elicit exemplar similarity effects that violated strict distance-from-boundary predictions. These effects can be characterized as instances in which the similarity of a test stimulus to a previously seen stimulus can cause violations of an explicit rule. The stimuli used in Brooks's experiments were imaginary animals whose features varied along five binary dimensions. Three of these five dimensions were relevant for categorizing the creatures as either "diggers" or "builders." If an animal had two of three builder features, the animal would be correctly classified as a builder; otherwise, it was a digger. Even when participants knew the rule, they were more likely to make errors if the most similar animal seen previously was from the opposite category. In this case, the most straightforward account of these data is provided by similarity-based, exemplar theories.

The frequency with which a particular stimulus is presented has also been shown to affect categorization performance. If one stimulus is presented more frequently than other stimuli, performance will be enhanced for that stimulus. It will be classified correctly more often and will be judged a more typical member of its category (Nosofsky, 1988a, 1988b, 1991a, 1991b; Nosofsky & Palmieri, 1997; Shin & Nosofsky, 1992). In this case also, exemplar-based theory provides the most parsimonious account of the graded category structure.

Models of Categorization

A number of different models have been established to formalize the principles of rule- and exemplar-based categorization. We focus on two of these. One tradition of exemplar-based categorization models, beginning with the context model (Medin & Schaffer, 1978) and leading through the generalized context model (Nosofsky, 1986, 1987, 1988b, 1989; Nosofsky, Clark, & Shin, 1989; Shin & Nosofsky, 1992) to ALCOVE (Choi, McDaniel, & Busemeyer, 1993; Kruschke, 1992, 1993a, 1993b, 1996b; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky & Kruschke, 1992; Nosofsky, Kruschke, & McKinley, 1992), has successfully accounted for a wide variety of classification phenomena.

Another tradition of rule-based categorization models with its genesis in the general recognition theory (GRT) has also been highly successful in accounting for a number of different phenomena (Ashby, 1988; Ashby & Gott, 1988; Ashby & Perrin, 1988; Ashby & Townsend, 1986). These models formalize various types of boundaries between categories, typically linear or quadratic in shape. A special case is linear boundaries orthogonal to the stimulus dimensions (e.g., Ashby, 1992; Nosofsky et al., 1989). We propose a hybrid model named ATRIUM[1] that combines these two traditions using the gating mechanism of Jacobs, Jordan, Nowlan, and Hinton (1991; see also Jacobs, 1997).

Goals of This Study

In this article, we describe two human categorization experiments designed to address three issues central to hybrid rule- and exemplar-based systems: the necessity of rules, the necessity of exemplar memory, and the interaction between these two subsystems in learning and in classification performance. We then highlight inadequacies in simple rule- or exemplar-based models of these behaviors, and we

[1] The name ATRIUM stands for Attention To Rules and Instances in a Unified Model.

describe ATRIUM and apply it to the experimental data. Both the empirical and modeling results suggest that human category learning is subserved by both rules and exemplars, which interact continuously.

HUMAN LEARNING EXPERIMENTS

The experiments in this study were designed to show the need for both rule- and exemplar-based models of categorization. Toward this end, three features of the category structures used in these experiments were key: (a) Some stimuli could be classified according to a rule, whereas other stimuli were exceptions and had to be memorized; (b) different training instances had different relative frequencies; and (c) some stimuli were never used in training and were available to examine generalization.

Experiment 1: Extrapolation Beyond Trained Instances

One essential element of human categorization behavior is *generalization*, or the ability to apply knowledge from past experience to novel situations. Two types of novel situations may be considered: those inside the range of training and those outside. The former instances are referred to as *interpolation* and the latter as *extrapolation*. A hypothetical classification task will help to elucidate this distinction. Imagine a category structure that follows a square-wave function. For example, stimuli in the range 50 to 59 would be assigned to Category A, stimuli in the range 60 to 69 to Category B, stimuli in the range 70 to 79 to category A again, and so forth (beyond both 50 and 79). If a finite number of stimuli are presented from a finite number of these regions—perhaps 40 stimuli selected randomly from four regions, extending from 50 to 89—then tests of novel stimuli between 50 and 89 are interpolative tests and tests of novel stimuli beyond this range are extrapolative tests. When presented with the task of interpolation, rule- and exemplar-based models of categorization will produce very similar results. When required to extrapolate beyond the training region, however, exemplar-based models of categorization cannot perform better than chance. If a rule-based model is able to induce the correct rule, it will continue to classify stimuli with the same degree of accuracy for extrapolation as for interpolation. Thus, extrapolative generalization is an important tool for distinguishing between rule- and exemplar-based generalization. A similar point was made by DeLosh, Busemeyer, and McDaniel (1997) for function learning.

In Experiment 1, we used stimuli consisting of a rectangle with a short interior line segment (an example is shown in Figure 1). These stimuli varied along two psychologically separable dimensions: rectangle height and the horizontal position of the line segment. These rectangles were shown with two accompanying scales marking values of rectangle height and line segment position from 0 to 9. From the training stimuli, participants could learn a one-dimensional rule that allowed them to classify most of the stimuli correctly. Figure 2 shows the category structure. Most of the training stimuli, the *regular* stimuli, could be classified according to a simple rule that divided the *primary* dimension at its midpoint (e.g., all rectangles taller than 4.5 are in one category; all those shorter are in the other). Two training stimuli were exceptions to the rule. These exceptions could only be identified accurately by attending to stimulus values on both the primary and the secondary dimensions. Each exception had its own category label, making four categories altogether.

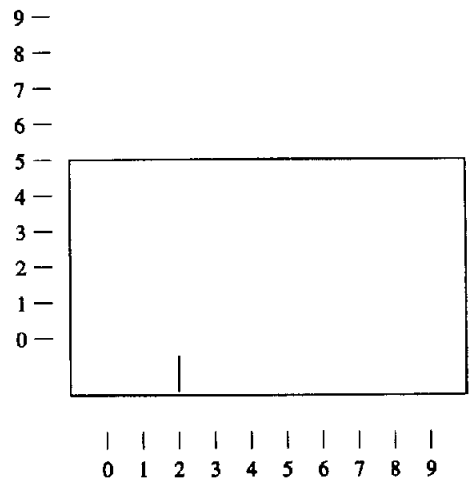

Figure 1. A sample stimulus from the experiments. On each trial, the rectangle and line segment as well as the numerical scales appeared on the screen. The rectangle height and line segment position were always aligned with 1 of the 10 values on the numerical scale.

All of the training stimuli had height and segment position values in the range of 1 to 8. All of the stimuli that were not presented during training were transfer stimuli used to test generalization. The four transfer stimuli labeled $\mathcal{T}_E$ and $\mathcal{T}_R$ in Figure 2 are important for two reasons: First, they require participants to extrapolate their category knowledge beyond the training region because they have extreme values on both dimensions of variation. Second, rule- and exemplar-based models make different predictions as to the pattern of responses participants should make to these $\mathcal{T}$ transfer stimuli.

To illustrate this difference, consider a case in which rectangle height is assigned as the primary dimension. A candidate set of rules that could correctly classify all the stimuli in this experiment would be: If the stimulus is the tall exception, classify it in the "tall exception" category. If the stimulus is the short exception, classify it in the "short exception" category. If these conditions were not met and the height exceeds 4.5, then classify it in the "tall" category. Otherwise classify it in the "short" category. A rule-based model such as this predicts no difference between the proportion of appropriate rule responses when the $\mathcal{T}_E$ stimuli are presented as opposed to when the $\mathcal{T}_R$ stimuli are presented.[2] An exemplar-based model, however, makes a different prediction. Because the $\mathcal{T}_E$ stimuli are most similar

[2] The rules just described might predict a difference if the $\mathcal{T}$ stimuli were easily confused with the training stimuli. The results from this experiment, however, suggest that this is not the case for human learners. For example, they are very accurate in their classifications along the rule boundary.

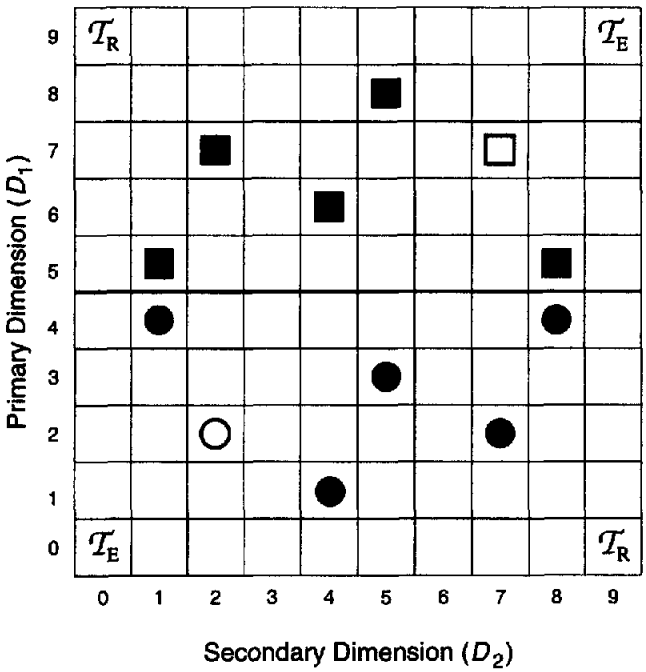

Figure 2. Category structure for Experiment 1. The rows and columns represent the stimulus values along each dimension (rectangle height or segment position). The cells containing filled shapes were rule training instances. Filled squares belong to one rule category and filled circles belong to another. The two cells containing open shapes were exception training instances. Each was the sole member of its exception category. The cells labeled $\mathcal{T}_E$ and $\mathcal{T}_R$ indicate the transfer stimuli used to distinguish between rule- and exemplar-based models of categorization.

to exception training instances, exemplar-based models predict that participants should give a higher proportion of exception responses to the $\mathcal{T}_E$ stimuli than to the $\mathcal{T}_R$ ones.

Method

Participants

The participants were 187 Indiana University undergraduate students drawn from introductory psychology classes. Of these, 41 were excluded from analysis because they did not meet the criterion of more than 50% correct in the last block of training. This criterion was chosen to select only those participants who had performed significantly better than chance as determined by probability matching. There were four different combinations of correct category label and frequency: Two occurred five times and two occurred twice in the last block of training. If participants probability matched, then the expected chance proportion correct over the last 14 trials was $p = 2 \cdot (5/14)^2 + 2 \cdot (9/14)^2 = .296$. Thus, using a binomial distribution with $p = .296$, $N = 14$, and a 95% confidence level, we arrived at the criterion of 50% (seven or more) correct. All participants were naive to experiments of this kind and received credit toward their final grade for participation.

Stimuli and Apparatus

The stimuli were rectangles that varied in height and contained a vertical line segment located near the base of the rectangle that varied its position (see Figure 1). The stimuli were presented with numerical scales so that each stimulus could be referenced by the corresponding scale values. The stimuli were presented on PC-compatible computers in individual, sound-dampened, dimly lit booths.

The category structure and training stimuli are shown in Figure 2. Each axis in Figure 2 represents one dimension of stimulus variation. Each cell represents a stimulus with the given values on each dimension. The training stimuli are indicated by circles and squares. The filled shapes indicate rule training stimuli. Filled squares belong to one rule category; filled circles belong to the other. The two open figures indicate exception training stimuli. Each was the sole member of its category. Thus, each stimulus was assigned to one of four categories. The transfer stimuli included the four $\mathcal{T}$ stimuli and every other untrained cell in Figure 2. To reduce the overall number of trials, each participant saw 50 of the 100 possible transfer stimuli. These 50 stimuli were all those that had even values (including 0) on the secondary dimension. The category structure was symmetrical about the rule boundary: Each half of the category structure, when rotated 180°, was identical with the other, unrotated half. After performing this rotation, even-numbered columns from one half of the structure match with odd-numbered ones from the other half. Thus, the transfer stimuli, after rotation, can be displayed in a 5-row × 10-column format, even though they were selected from only even columns during the experiment.

The category structure was counterbalanced between participants by using its horizontal or vertical mirror image or by assigning either of the two dimensions of stimulus variation (i.e., rectangle height or segment position) to the primary dimension and the other to the secondary dimension. This yielded eight different physical realizations of the abstract structure. Because of this counterbalancing, every possible physical stimulus was presented during the transfer block across the course of counterbalancing. Because of the symmetry of the abstract structure, every possible abstract stimulus was also presented during the transfer block.

Procedure

Participants were trained over the course of 29 blocks of 14 trials each. At the end of every third block, participants were given a self-timed rest period. Within each block, each of the rule training stimuli shown in Figure 2 was presented once. The exception training stimuli were each presented twice per block.

In each training trial, a stimulus was presented and participants were instructed to assign it to one of four categories by pressing one of the computer keys—*S*, *F*, *J*, or *L*—as quickly as possible without making errors. When a response was made or the response period ended, feedback was given. Participants were told whether their selection was right or wrong. If their selection was wrong, the computer generated a tone to signal their error. If they did not respond within 6 s, the computer generated a high-pitched tone and displayed "Faster!" Then the correct answer was displayed for 1 s.

After the training blocks concluded, participants were told that they were to assign labels to the rectangles as before. They were told, however, that rectangles they had not seen previously would be shown, that they should make their best guess, and that they would not receive any feedback. During this block of trials, the transfer stimuli were displayed in a random order.

Results

Before performing other analyses, we examined participants' responses to see whether the between-participants counterbalancing of dimensions had any significant influence on performance. In postexperiment interviews, a num-

ber of participants reported that they had noticed that the exception stimuli (the open shapes in Figure 2) occurred when the value of rectangle height was equal to the value of the segment position. (Recall that scales were available below and next to the rectangle for reference.) Thus, rather than using the rules described previously (and in the case that height is assigned to the primary dimension), these participants may have been using rules like: If the rectangle height is the same as the line segment position and the height exceeds 4.5, classify it in the "tall exception" category. If the rectangle height is the same as the line segment position and the height does not exceeds 4.5, classify it in the "short exception" category. If these conditions were not met and the height exceeds 4.5, then classify it in the "tall" category. Otherwise, classify it in the "short" category. This may be thought of as an "equal-value abstraction" for describing exceptions to the primary rule. Whereas the notion of an "exception" can thereby be extended from a single stimulus that violates a rule to a defined set of rule-violating stimuli, the category structure we intended participants to induce consisted of two categories that could be distinguished by a unidimensional rule and two exception categories that contained one stimulus each. Our goal was limited to examining how people use exemplar-based representation and a single rule. Because it is beyond the scope of this article to address the use of multiple, rulelike abstractions, we have excluded data from conditions in which this unintended solution was available. We discuss this further in the General Discussion.

The equal-value abstraction was available in four of the eight counterbalanced conditions; in the others, the simplest equivalent abstraction for exceptions is a "sum to nine" abstraction. That is, if the sum of the rectangle height and segment position is nine, then the stimulus is classified as an exception. We compared generalization performance in the conditions in which the equal-value abstraction was available with the conditions in which the sum-to-nine abstraction was available to determine whether these abstractions were induced with equal probability. For each participant, we computed the difference between the proportion of exception responses for stimuli that fulfilled the exception abstraction and the proportion of exception responses for stimuli that failed to fulfill the abstraction. The participants in the four conditions that yielded the equal-value abstraction gave more exception responses for stimuli meeting its conditions ($M = .15$, $SD = .36$) than did those in the conditions that yielded the sum-to-nine abstraction ($M = .01$, $SD = .26$), $t(144) = 2.55$, $p = .01$ (see also Figure A2).

The results from the conditions that were more likely to yield the intended category structure ($N = 62$) are reported in the main text. Results from the other conditions are described in Appendix A.

Training

Although the focus of this experiment is participants' responses to novel stimuli during the transfer phase of the experiment, performance during training is important for two reasons: First, for the transfer data to be meaningful, the participants must have learned the training stimuli. Second, patterns of performance during learning might imply use of either a rule- or an exemplar-based strategy.

Correct responses to rule training stimuli (the filled shapes in Figure 2) showed improvement from 25% correct (chance) to 89% correct (see the left panel in Figure 3), whereas responses indicating the exception to the rule (i.e., open shape responses to the corresponding filled shape, referred to as "near-exception" responses) decreased from chance to 3%. A similar analysis of responses to the exception training stimuli (the open shapes in

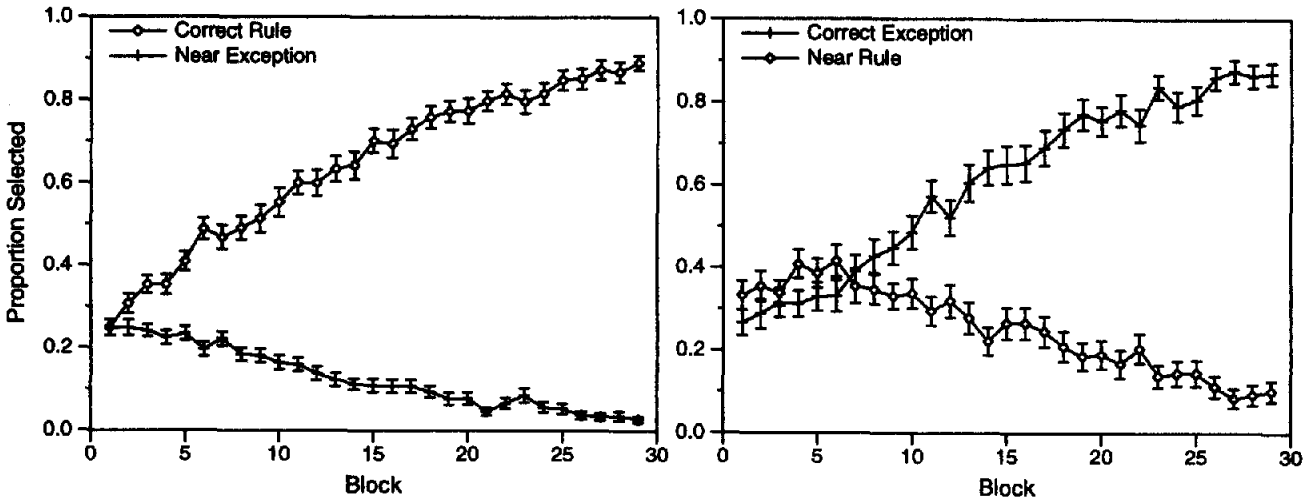

Figure 3. The left panel shows the proportion of correct rule responses and near-exception responses by block in Experiment 1. The right panel shows the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 1. In both panels, error bars extend 1 *SE* above and below the mean.

Figure 2) shows the same pattern of improvement from chance to 87% correct (see the right panel of Figure 3).

The phenomenon of participants classifying an instance of an exception category with a label appropriate for the surrounding rule stimuli is referred to as *overgeneralization*. The proportion of participants' responses indicating the correct rule category for stimuli surrounding the exception (i.e., filled shape responses to the corresponding open shape, referred to as "near-rule" responses), however, starts above chance (at 33%), remains greater than the proportion of correct responses until after Block 6,[3] and continues greater than chance through Block 10 ($M = .34$, $SD = .28$), $t(61) = 2.45$, $p = .02$. Some authors (e.g., Ervin, 1964) have used over-generalization as evidence that a rule has been induced (cf. Rumelhart & McClelland, 1986). In this instance, however, we advise caution in drawing such a conclusion. In this category structure, as we show later, the similarity between the rule training instances and the exception training instances is sufficient for this phenomenon to be accounted for by an exemplar-based model with no mechanism for rule induction.

During training, then, the participants did learn to classify the training instances correctly. Although they showed a pattern of overgeneralization, this pattern does not distinguish rule-from exemplar-based categorization.

Transfer

The principle purpose for the transfer task was to compare the proportion of participants' exception responses to the $\mathcal{T}_R$ ($M = .10$, $SD = 0.30$) and $\mathcal{T}_E$ ($M = .11$, $SD = 0.32$) stimuli. As is predicted by the rule-based categorization strategy described previously, no significant difference was found between the two proportions, $t(61) = 0.30$, $p = .77$.[4] Although we chose to compare participants' responses to the $\mathcal{T}_E$ and $\mathcal{T}_R$ stimuli because these stimuli were beyond the range of training on both stimulus dimensions, we also compared participants' responses for the stimuli that were adjacent to the $\mathcal{T}$ stimuli (not including the one on the diagonal) to show the generality of this phenomenon. Even considering these stimuli, participants did not give a significantly greater proportion of exception responses to stimuli adjacent to the $\mathcal{T}_E$ stimuli ($M = .13$, $SD = 0.27$) than to the stimuli adjacent to the $\mathcal{T}_R$ stimuli ($M = .10$, $SD = 0.21$), $t(61) = 0.62$, $p = .54$. Though these data are consistent with the rule-based account of categorization described previously and seem to contradict the predictions of an exemplar-based model, these results are merely suggestive until candidate models are actually fit to the data. Nevertheless, much as in the experiment by Rips (1989), participants in this experiment seem to be classifying these stimuli contrary to what would be anticipated by an exemplar-based model. When participants extrapolate their category knowledge, they appear to be able to use rule- rather than exemplar-based representation.

Although these data are suggestive of rule-based processing, the pattern of classification for stimuli near the exception training stimulus also shows hallmarks of exemplar-based classification. Chief among these is simply that as the similarity between the exception training stimulus and the transfer stimuli decreases, the proportion of exception responses decreases.

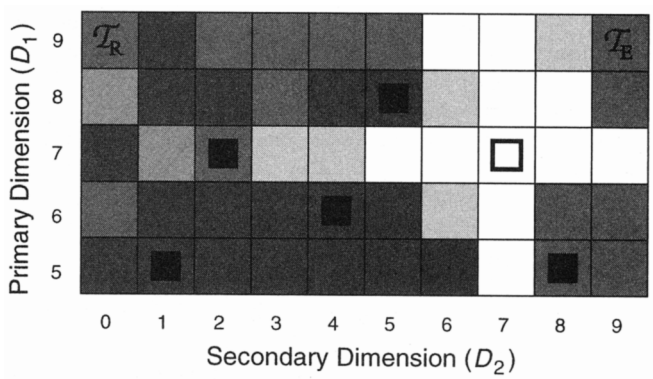

Figure 4. Proportion of exception responses in the transfer phase of Experiment 1. The shading in each cell indicates the proportion of exception responses. Light cells indicate a high proportion of exception responses; dark cells indicate a low proportion of exception responses. This diagram shows the top half of the category structure for Experiment 1. Data from the bottom half have been rotated and combined with those in the top half to generate this diagram. Training instances are marked with a filled or an open square (rule or exception, respectively), and the test stimuli described in the text are marked with $\mathcal{T}_R$ or $\mathcal{T}_E$.

Further examination of the exception responses illustrates interesting details about the nature of the exemplar-based categorization in this task. Nosofsky (1984) showed that participants allocate attention to each stimulus dimension in such a way as to improve their performance. If the rule training instances affected participants' exemplar-based categorization in this task, participants would be expected to attend to the primary dimension more than the secondary dimension. This means that a change in the primary dimension would be more noticeable than a change of equal size in the secondary dimension (Goldstone, 1994). This would be reflected in the results shown in Figure 4 by a higher proportion of exception responses for stimuli that matched the exception on the primary dimension (i.e., $D_1 = 2$ or $D_1 = 7$, excluding the rule training instances in Figure 2) than for those that matched on the secondary dimension (i.e., $D_2 = 2$ or $D_2 = 7$, excluding the rule training instances in Figure 2). In this experiment, participants did tend to give more exception responses when presented with stimuli that matched the exception on the primary dimension ($M = .22$, $SD = .23$) than when presented with stimuli that matched on the secondary dimension ($M = .14$, $SD = .18$), although this difference was only marginally significant, $t(61) = 1.85$, $p = .07$.[5]

[3] At no point, however, is the proportion of rule responses significantly greater than the proportion of exception responses.

[4] A power analysis indicated that, for the variance obtained, a mean difference of .10 between the two proportions would yield a power of .60.

[5] This difference cannot be attributed to difference in the salience of the stimulus dimensions because these were counterbalanced.

This trend toward rule-appropriate dimensional attention in exception classification can be interpreted as an interaction between rule- and exemplar-based classification. An assumption underlying many hybrid systems is that each subsystem is used only in those cases for which it is best suited. In this case, it might be assumed that rule training instances are fed through a rule subsystem and exceptions are fed through an exemplar subsystem. If, however, these two systems had this degree of independence, the exemplar subsystem would be expected to allocate attention optimally for the exceptions alone and to ignore the influence of the rule. Because proper exception classification requires attention to both stimulus dimensions, allocating attention appropriately for the classification of rule training instances is detrimental to exception classification. Nevertheless, exemplar-based classification of stimuli near the exceptions shows influences of the rule-learning process.

The results of Experiment 1, then, show evidence of rule- and exemplar-based categorization. Rule-based categorization was demonstrated by examining participants' extrapolation, and exemplar-based categorization was demonstrated by examination of stimuli that were similar to the exception training exemplar. In addition, we showed evidence of overgeneralization in learning and exemplar-based representation with dimensional attention. Hence, besides showing the necessity of rule- and exemplar-based categorization, we further showed that the rule structure influenced participants' dimensional attention in exemplar-based categorization even when it was not necessarily advantageous for classification of the exception.

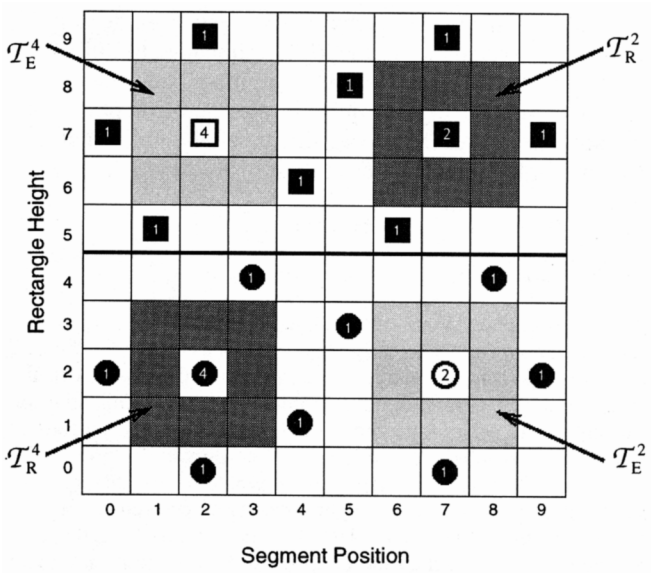

Figure 5. Category structure for Experiment 2. The rows and columns represent the stimulus values along each dimension (rectangle height and segment position). Each training stimulus is designated by a cell containing a filled or open shape; a filled shape denotes a rule training instance and an open shape denotes an exception. The filled squares belong to one rule category and filled circles belong to the other. The open shapes were exception training instances: Each was the sole member of its exception category. The numbers in each shape indicate the relative frequency of that training stimulus. The regions labeled $\mathcal{T}_E$ and $\mathcal{T}_R$ indicate transfer stimuli used to compare the influence of manipulating the frequency of rule and exception training instances.

Experiment 2: Training Instance Frequency Effects

Recall that one emphasis of this research is to examine the interactions within a hybrid rule- and exemplar-based categorization system. In Experiment 1, we showed that the category structure of the rules influenced transfer performance for $\mathcal{T}$ stimuli that were similar to the exceptions. To explore this interaction further in Experiment 2, we manipulated the frequency with which both rule and exception training exemplars were presented.

A strict rule-based model predicts that changing the frequency of rule training instances would have no effect on generalization once the rule boundary is correctly situated. The boundary would divide the two categories at the appropriate point without any memory for specific rule training instances. For example, the same set of rules that were described in Experiment 1 would correctly classify all the training stimuli in this experiment (Figure 5). Likewise, a hybrid model that assumes rule training instances are categorized by the rule module and exception training instances are categorized by the exemplar module would predict no influence of the frequency of rule training instances once the correct rule boundary had been established. If the frequency of rule training instances does influence performance throughout training, this interaction would provide an important constraint for a hybrid model of categorization. It would require exemplar memory of rule training instances as well as of exceptions.

There has been some dispute about the effects of instance training frequency on categorization performance. Whereas Homa, Dunbar, and Nohre (1991) found that manipulations of training instance frequency were relatively ineffective when learning had proceeded to a criterion of near-perfect performance, Nosofsky (1988a, 1988b, 1991a, 1991b; Nosofsky & Palmieri, 1997; Shin & Nosofsky, 1992) found a robust influence of training instance frequency, even for categories that may be specified by a rule. Rather than testing generalization only at the end of training, when the effect of training instance frequency may be difficult to detect, we examined generalization to untrained stimuli throughout the course of learning.

Many aspects of Experiment 2 were similar to Experiment 1. The category structure in this experiment allowed most stimuli to be correctly categorized using a rule that divided a single stimulus dimension into two parts, and two stimuli served as exceptions to this rule, as shown in Figure 5. To categorize the exceptions correctly, it was necessary to attend to both stimulus dimensions. As before, the stimuli were rectangles that varied in height with vertical line segments near the bottom of the rectangles that varied in their horizontal position (see Figure 1 for an example).

To explore the effect of training instance frequency, we manipulated the presentation frequency of two of the rule training instances and both of the exception training in-

stances. We measured participants' responses to stimuli immediately surrounding the high-frequency stimuli (those labeled $\mathcal{T}$ in Figure 5). By examining the pattern of generalization for high-frequency rules and exceptions, we are able to constrain further the type of model that can account for human behavior in these types of tasks.

Method

Participants

The participants were 109 Indiana University undergraduate students drawn from introductory psychology classes. Of these, 7 were excluded from analysis because they did not meet the criterion of 50% correct in the last block of training. This criterion was chosen to select only those participants who had performed significantly better than chance as determined by probability matching. There were six different combinations of category label and frequency: two occurred eight times, two occurred four times, and two occurred twice in the last block of training. If participants probability matched, then the expected chance proportion correct over the last 28 trials was $p = 2 \cdot (8/28)^2 + 2 \cdot (4/28)^2 + 2 \cdot (2/28)^2 = .214$. Thus, using a normal approximation to a binomial distribution with $p = .214$, $N = 28$, and a 95% confidence level, we arrived at a criterion of 50% (14 or more) correct. All participants were naive to experiments of this kind and received credit toward their final grade for participation.

Stimuli and Apparatus

The stimuli were the same as those used in Experiment 1. The category structure and training stimuli are shown in Figure 5. The vertical axis represents the height of the rectangle, and the horizontal axis represents the position of the line segment. Each cell in Figure 5 represents a stimulus with given values on each dimension. The training stimuli are indicated by cells containing shapes. The filled shapes signify rule training instances: Squares represent one rule category and circles represent the other. The open shapes signify exception training instances: The square represents a one-member exception category and the circle represents another. The numbers inside the shapes represent the relative frequency of the stimuli. All of the untrained stimuli represented in Figure 5 were used to test generalization.

The category structure was counterbalanced between participants by using its horizontal or vertical mirror image. We also counterbalanced the frequencies of the high-frequency stimuli (those stimuli that occurred two or four times per block). As shown in Figure 5, the Frequency 4 exception and the Frequency 2 rule training stimuli are both in the square rule region. We refer to this as the *mixed-frequency* condition. This condition was counterbalanced with one in which the Frequency 4 exception and the Frequency 4 rule training stimuli were in the same rule region. We refer to this as the *same-frequency* condition. This yielded eight different category structure conditions. We did not counterbalance the assignment of the primary and secondary dimensions to the two physical dimensions of variation, because analysis from Experiment 1 showed no behavioral difference between the two conditions. Thus, the primary dimension was always assigned to height, and the secondary dimension was always assigned to line segment position.

Procedure

Participants were trained over the course of 16 blocks of 28 trials each. Within each block, each of the training stimuli shown in Figure 5 was presented one, two, or four times according to the number in the corresponding shape. Each training trial proceeded in the same way as in Experiment 1.

Because of our concern about ceiling effects at the end of training, participants were presented with a block of 14 transfer trials after each training block. Before the transfer trials began, participants were told that they were to assign labels to the rectangles as before. They were told, however, that rectangles they had not seen previously would be shown, that they should make their best guess, and that they would not receive any feedback. For each block of these trials, 14 of the 100 possible stimuli were randomly selected and displayed.

Results

As in Experiment 1, we examined participants' responses to see whether their patterns of generalization varied as a function of the counterbalancing. Specifically, we looked for differences in generalization depending on whether participants were in an "equal-value" or "sum-to-nine" exception condition. Here, however, we limited the analysis to the final eight blocks inasmuch as participants were most likely to be using abstractions in these blocks. In this case, as before, the participants in the conditions that yielded the equal-value abstraction gave more exception responses for stimuli meeting its conditions ($M = .04$, $SD = .12$) than did those in the conditions that yielded the sum-to-nine abstraction ($M = 0.01$, $SD = .03$), $t(100) = 2.55$, $p = .01$.

Because participants who are using an exception abstraction are not viewing the exceptions as individual anomalies, we only present the results from conditions that were more likely to prompt the intended category structure ($N = 47$). The results from the other conditions are presented in Appendix A.

Training

As in Experiment 1, we analyze the training data for (a) evidence that the category structure was learned and (b) evidence of overgeneralization.

As shown in Figure 6, participants' rule and exception classification performance improved over the course of training. When classifying rule training instances, they improved from 37% correct in Block 1 to 94% correct in Block 16 (left panel of Figure 6). Likewise, their performance classifying exception training instances improved from 23% correct in Block 1 to 82% correct in Block 16 (right panel of Figure 6).

As the right panel of Figure 6 shows, participants overgeneralized extensively throughout the first several blocks. A comparison of the proportion of participants' rule and exception responses when an exception training stimulus was presented shows overgeneralization through Block 4 ($M = .21$, $SD = .53$), $t(46) = 2.83$, $p = .007$, and a comparison of the proportion of participants' rule responses relative to chance (.25) shows overgeneralization through Block 7 ($M = .33$, $SD = .26$), $t(46) = 2.11$, $p = .04$.

Over the course of all the training trials, participants performed better when classifying stimuli that appeared four times per block than when classifying those that appeared

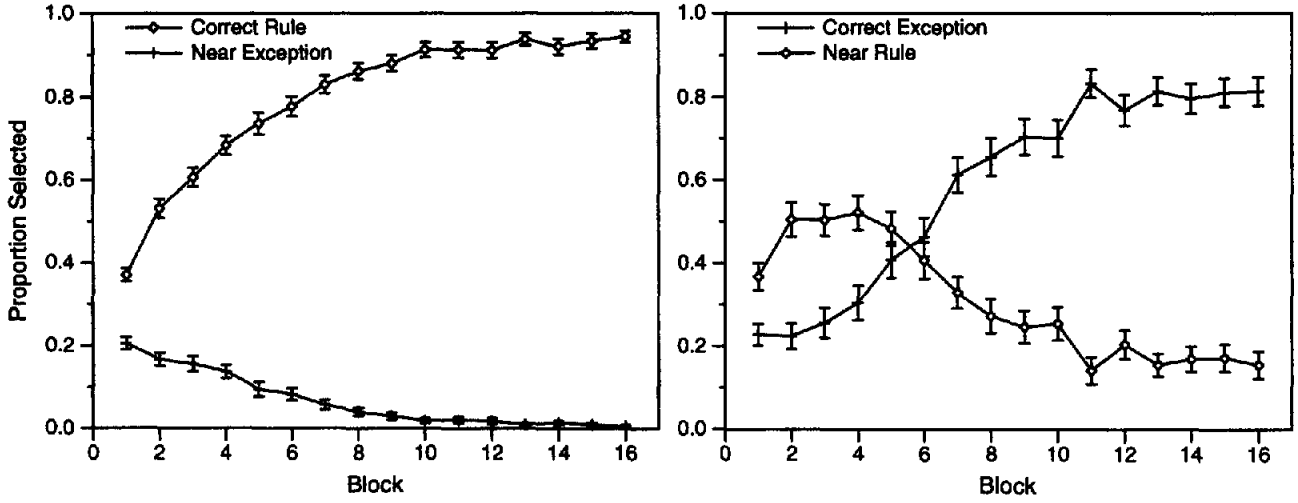

Figure 6. The left panel shows the proportion of correct rule responses and near-exception responses by block in Experiment 2. The right panel shows the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 2. In both panels, error bars extend 1 *SE* above and below the mean.

twice per block, 75% versus 66% correct, $F(1, 46) = 36.99$, $MSE = 0.0619$, $p < .0001$.[6] Performance was reliably enhanced by presentation frequency for both rule and exception training instances. The mean difference in the proportion of correct responses between the Frequency 4 and Frequency 2 rule training instances was .06 ($SD = .13$), $t(46) = 3.50$, $p = .001$, and the mean difference between the Frequency 4 and Frequency 2 exception training instances was .12 ($SD = .14$), $t(46) = 5.5639$, $p < .0001$. The facilitation for these exception instances, however, was not reliably greater than facilitation for Frequency 4 over the Frequency 2 rule training instances $F(1, 46) = 0.97$, $MSE = 0.0544$, $p = .33$. This influence of presentation frequency for rule training instance contradicts a strict rule-based interpretation or a rule-plus-exemplar interpretation that limits exemplar representation exclusively to exceptions. If the rule training instances were classified using only rule-based representation, there would be no memory for any single rule training instance. Rather, representation would consist of a boundary separating high- and low-frequency instances alike.

During training, then, participants learned the intended category structure after initial overgeneralization. Participants learned the Frequency 4 training instances more quickly than the Frequency 2 training instances for both rule and exception training instances.

Transfer

Those theories that predict that variation of rule training instance frequency will have no effect on learning also predict no effect on generalization to nearby stimuli. To test this prediction, we calculated the proportion of rule-based responses for the stimuli in the shaded areas labeled $\mathcal{T}$ in Figure 5. (The complete set of rule-response proportions is given in Appendix B.) Each of the $\mathcal{T}$ cells was adjacent to a rule or exception training stimulus that was presented two or four times per block. The type and frequency of the training stimulus are indicated by the $\mathcal{T}$ subscript and superscript. Figure 7 shows the average proportion of appropriate rule responses for each set of $\mathcal{T}$ stimuli. As anticipated, the $\mathcal{T}_E^4$ stimuli showed fewer rule responses ($M = .73$, $SD = .21$) than did the $\mathcal{T}_E^2$ stimuli ($M = .79$, $SD = .20$). The mean of the arcsine transformed differences was 0.14 ($SD = 0.41$), $t(46) = 2.39$, $p = .02$. If, however, exemplar information was being used for high-frequency rules as well as for exceptions, the key test would be to show an influence of *rule* training instance frequency. This test likewise showed an effect of training instance frequency: Participants gave *more* rule responses to the $\mathcal{T}_R^4$ stimuli ($M = .87$, $SD = .13$) than they did to the $\mathcal{T}_R^2$ stimuli ($M = .81$, $SD = .21$). The mean of these arcsine transformed differences was 0.21 ($SD = 0.47$), $t(46) = 3.03$, $p = .004$. In terms of Figure 7, this means that the slope of the solid line is significantly greater than zero.

An alternative explanation of these results, however, might be that because the proportion of rule responses for the $\mathcal{T}$ stimuli were collapsed across all 16 transfer blocks, these results might merely be an effect of the participants' placement of the rule boundary in the early training trials. If this were the case, one would expect that any difference in the proportion of rule responses for the $\mathcal{T}_R^4$ and $\mathcal{T}_R^2$ stimuli would disappear by the end of learning. Figure 8 shows the proportion of rule responses for the $\mathcal{T}_R^4$ and $\mathcal{T}_R^2$ stimuli for each block of transfer trials. It appears that performance reached asymptote at Block 12. Even after reaching asymptote, however, participants give significantly more rule responses when presented with the $\mathcal{T}_R^4$ stimuli ($M =$

[6] Although the numbers presented throughout this section are percentage (or proportion) correct, the dependent measure for the statistical analyses was an arcsine transformation of proportion correct to meet assumptions of normality better.

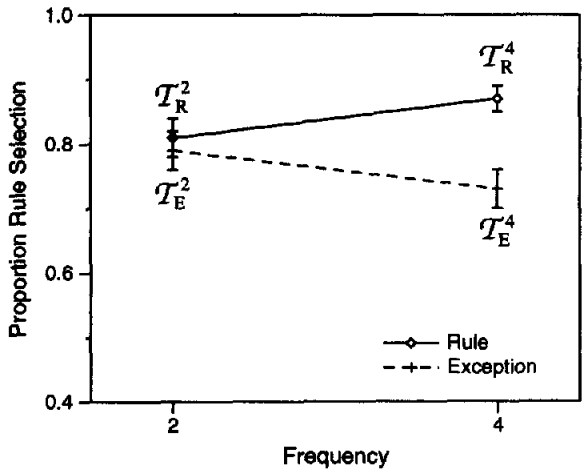

Figure 7. The proportion of appropriate rule responses for the $\mathcal{T}$ stimuli in Experiment 2.

.96, $SD = .11$) than when presented with the $\mathcal{T}_R^2$ stimuli ($M = .91$, $SD = .19$). The mean of the arcsine transformed differences was 0.14 ($SD = 0.37$), $t(46) = 2.79$, $p = .008$. Thus, even after participants' performance had stabilized, rule generalization was stronger for the $\mathcal{T}_R^4$ stimuli than for the $\mathcal{T}_R^2$ stimuli.

To test whether the improved rule generalization for the $\mathcal{T}_R^4$ stimuli might have been due to an overall advantage for this rule category rather than localized exemplar memory, we tested generalization to the remaining untrained stimuli. In Blocks 12–16, participants did not give significantly more rule responses for untrained stimuli in the same half of the stimulus space as the Frequency 4 rule training instance, excluding the $\mathcal{T}_R^4$ stimuli ($M = .89$, $SD = .17$), than for the untrained stimuli in the same half of the stimulus space as the Frequency 2 rule training instance, excluding the $\mathcal{T}_R^2$ stimuli, ($M = .89$, $SD = .16$). The mean arcsine transformed difference was 0.0002 ($SD = 0.37$), $t(46) = 0.004$, $p = .997$. These results, then, support the hypothesis that the increased proportion of rule responses for the $\mathcal{T}_R^4$ stimuli versus the $\mathcal{T}_R^2$ stimuli was due to exemplar memory for the high-frequency rule training stimuli rather than a general improvement for one rule category over the other.

Discussion

The goals of these experiments were to show that (a) exemplar representation alone is insufficient to account for aspects of human classification behavior and (b) rule representation is insufficient to account for patterns of classification, even for stimuli that can be classified according to a rule.

In Experiment 1, participants showed a pattern of classification that violated predictions based on similarity to memorized exemplars. Whereas the exception test stimuli ($\mathcal{T}_E$) were more similar to the exception training stimuli than to any rule training stimuli, participants classified them according to the rule at the same rate as the contrasting rule test stimuli ($\mathcal{T}_R$).

One concern that may be raised about this conclusion is that, as explained previously, exemplar-based models can selectively allocate attention to the component stimulus dimensions to optimize performance (Nosofsky, 1984). Because differential weighting improves performance by changing the distance relations between various stimuli, one might hypothesize that an exemplar model could, indeed, account for these data. This concern can only be fully answered by formally modeling exemplar-based categorization.

Although Experiment 1 provided evidence for rule-based representation, simple, all-or-none rule-based representation by itself cannot account for the complete pattern of results obtained. The most important of these is the pattern of exception responses for stimuli similar to the exception training instances. As similarity to the exception training instances decreased, participants' exception generalization also decreased, as predicted by exemplar theories (Nosofsky, 1984; Shepard, 1987).

In addition to providing evidence of both rule- and exemplar-based categorization, the results from Experiment 1 suggest that participants are shifting attention to the primary dimension, even when classifying exceptions. This implies that the exemplar system cannot be considered an exception system. If the only role of exemplar memory were to classify exceptions, it should allocate attention to maximize performance in this task alone. The evidence of an attentional shift appropriate to rule classification indicates that the exemplar system is processing information for both rule and exception training instances.

In Experiment 1, the influence of the rule training instances on the exemplar system had a deleterious effect on exception performance. These results led to the question: In what cases might this influence be helpful rather than detrimental? Experiment 2 showed that the exemplar system can serve to augment rule-based classification by learning associations between highly salient rule training instances and their correct category assignment.

Participants' responses in Experiment 2 showed that rule training instance frequency affected classification perfor-

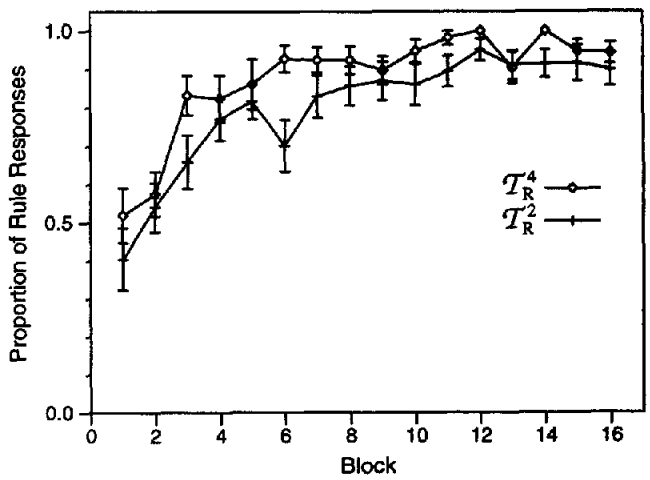

Figure 8. The proportion of appropriate rule responses for the $\mathcal{T}_R$ stimuli in Experiment 2 by block.

mance in much the same way as exception training instance frequency. This would not be possible if the rule instances were not represented in memory. If rule instances were consigned to rule-based representation alone, no information about specific instances would be retained; hence, it would not be possible to represent their relative frequency.

These results, then, suggest that categorization involves both rule- and exemplar-based representation. Also, exemplar-based representation appears to learn instances of rules as well as instances of exceptions.

ATRIUM: A HYBRID CONNECTIONIST MODEL

Evidence from human learning therefore suggests that rule- and exemplar-based representations are both necessary to capture certain aspects of performance. We first outline ATRIUM, a hybrid model that incorporates rule- and exemplar-based representation. We then evaluate the performance of the exemplar-based portion of the model alone, followed by an evaluation of the full model.

ATRIUM, which was first described by Kruschke and Erickson (1994), is composed of a rule module, an exemplar module, and a competitive gating mechanism that links the two modules together. The general architecture of the model is shown in Figure 9.

Every stimulus presented to the model is processed simultaneously by the rule module and the exemplar module. The rule nodes within the rule module are activated according to where the stimulus falls relative to the rule boundary. One node is activated if the stimulus falls on one side of the boundary; the other node is activated if it falls on the other side. Each rule node is connected to all the rule category nodes by learned, weighted connections. Thus, the rule module learns to associate rule-bounded regions of psychological space with category selections.

The exemplar module receives the same input as the rule module, but processes it as in the ALCOVE model described by Kruschke (1992). The input is interpreted as a point in psychological space that activates nearby exemplar nodes strongly and distant exemplar nodes more weakly. Each exemplar node is connected to all the exemplar category nodes by learned, weighted connections.

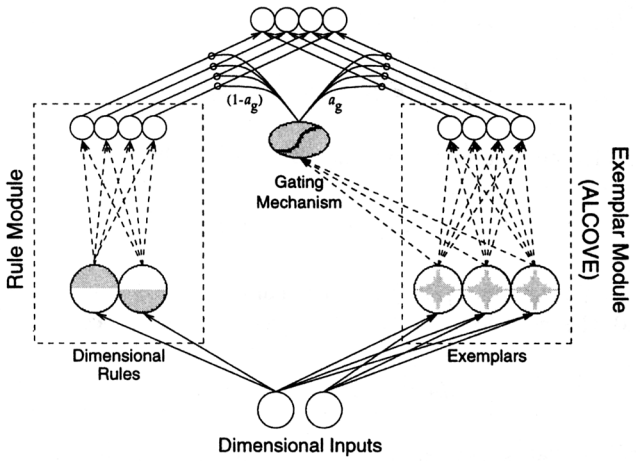

Figure 9. The architecture of ATRIUM, the hybrid rule and exemplar model, used to fit the experimental data. The dotted lines represent connections with learned weights.

The final output is a combination of the rule module output and the exemplar module output. The degree to which each module contributes is determined by the output of the gate node. The final output probabilities are generated by normalizing the activation of each output category node relative to the total activation of all the output category nodes.

The gating mechanism receives input from the exemplar nodes via learned, weighted connections. The output of the gate node is a sigmoidal function of the sum of these weighted inputs. As this sum increases, the likelihood that the final output of the model is based on the exemplar category node activations increases. The connections between the exemplar nodes and the gate node learn the degree to which regions in psychological space are best classified by each module.

Rule Module

The rule module contains two types of nodes: rule nodes and category nodes. The rule nodes implement a linear sigmoid that is aligned with the primary dimension and has its bias adjusted so that the midpoint of the sigmoid falls on the rule boundary (see Equation 1). This is equivalent to using a step function at the rule boundary with the addition of normally distributed perceptual or criterial noise. We used nodes that respond optimally to explore how well a gated hybrid system, including a rule module, could work in this task. Moreover, the results from the transfer trials in our experiments showed that induction of such a rule is plausible.

The activation of the rule nodes depends only on the rule dimension. One node responds to large values and the other to small, as follows:

$$a_{\text{small}} = 1 - [1 + \exp[-\gamma_r(D_1 + \beta_1)]]^{-1}$$
$$a_{\text{large}} = [1 + \exp[-\gamma_r(D_1 + \beta_1)]]^{-1}. \quad (1)$$

Here, γ_r is the gain of the sigmoid corresponding to a function of the standard deviation of the normally distributed noise, D_1 represents the stimulus value on the primary dimension, and β_1 represents the bias as described previously. These two rule nodes both have weighted connections to each of the four rule-module category nodes, one for each possible category. The activation a_{r_k} of rule-module Category Node k is given by

$$a_{r_k} = w_{r_k,\text{large}} a_{\text{large}} + w_{r_k,\text{small}} a_{\text{small}}, \quad (2)$$

where $w_{r_k,\text{large}}$ is the connection weight from the large-value rule node to rule-module Category Node k and $w_{r_k,\text{small}}$ is the connection weight from the small-value rule node to rule-module Category Node k.

Exemplar Module

The exemplar module is a full implementation of ALCOVE (Kruschke, 1992). The input to the exemplar module is the same

as that to the rule module. It is interpreted as a point in psychological space that activates nearby exemplar nodes strongly and distant exemplar nodes more weakly. Each exemplar node is connected to all the exemplar category nodes by learned, weighted connections that represent the association between each exemplar and each category.

Let the position of Exemplar Node j be represented by $(h_{e_{j1}}, h_{e_{j2}}, \ldots)$. Then the activation a_{e_j} of Exemplar Node j is expressed

$$a_{e_j} = \exp\left[-.5c\sum_i \alpha_i |h_{e_{ji}} - d_i|\right], \tag{3}$$

where c is the specificity of the node, α_i is the dimensional attention strength for Dimension i, and d_i is the coordinate of the stimulus on Dimension i. One hundred exemplar nodes were positioned so that their segment position and rectangle height values were located at psychological values obtained in a separate scaling study described in Appendix C.

The activation of each of the four exemplar-module category nodes is obtained as a weighted sum of all the exemplar node activations,

$$a_{e_k} = \sum_{e_j} w_{e_k e_j} a_{e_j}, \tag{4}$$

where w_{e_k,e_j} is the connection weight from Exemplar Node j to exemplar-module Category Node k.

Gating Mechanism

The gating node serves to pass a proportion of the activation from both sets of category nodes to a final "output" set of category nodes. The proportion is governed by the gating node's activation:

$$a_g = \left\{1 + \exp\left[-\gamma_g \sum_{e_j} w_{g,e_j} a_{e_j} + \beta_g\right]\right\}^{-1}, \tag{5}$$

where w_{g,e_j} is the connection weight from Exemplar Node j to the gating node, β_g is the gate bias, and γ_g is the gate gain. The activation of the gate node is squashed in the range (0, 1) to represent the probability of using the exemplar module. This probability is a function of the activation of each of the exemplar nodes and the learned weights connecting those nodes with the gating node. This allows the model to learn which module is best suited for particular exemplars.

The probability of choosing Category K, the mixed-module choice probability, is computed as follows:

$$p(K) = a_g \frac{\exp(\phi a_{e_K})}{\sum_k \exp(\phi a_{e_k})} + (1 - a_g)\frac{\exp(\phi a_{r_K})}{\sum_k \exp(\phi a_{r_k})}, \tag{6}$$

where ϕ is a scaling constant, which may be thought of as representing the level of "decisiveness" in the system. If ϕ is low, differences in activation are diminished in the final output probabilities; if ϕ is high, differences in activation are accentuated.

As described and as used for simulations, ATRIUM is deterministic. The gating mechanism described by Jacobs et al. (1991), however, is stochastic. In their formulation, a_g does not weight the category predictions from each module; it is the probability that a given module is used, and hence, only one module is actually selected on each trial. A version of ATRIUM implemented with this stochastic mechanism should exhibit the same average behavior as the deterministic version described here.

Learning

Learning is achieved by gradient descent on error. The error is computed using an adaptation of Equation 1.3 from Jacobs et al. (1991) combined with humble teachers as defined by Kruschke (1992, 1996a). Let $\mathbf{t}_m$ be a vector of humble teacher values such that

$$\mathbf{t}_{m_k} = \begin{cases} \tilde{1} = \max(1, a_{m_k}) & \text{if } k \text{ is correct} \\ \tilde{0} = \min(0, a_{m_k}) & \text{otherwise,} \end{cases} \tag{7}$$

and $\mathbf{a}_m$ be the output vector of Module m, where m is either r or e for rule or exemplar module, respectively. The error, then, is

$$\begin{aligned} E &= -\log\left\{\sum_{\text{mod } m} p(m) \exp\left[-.5c_m \|\mathbf{t}_m - \mathbf{a}_m\|^2\right]\right\} \\ &= -\log\left\{\sum_{\text{mod } m} p(m) \exp\left[-.5c_m \sum_{\text{cat } k} (t_{m_k} - a_{m_k})^2\right]\right\}, \end{aligned} \tag{8}$$

where $c_m \geq 0$ is the "cost" of Module m, and p(m) is the probability that Module m is selected. The probability of selecting the exemplar module is a_g. Let the accuracy of the rule module be defined as

$$RA = \exp(-.5c_r \|\mathbf{t}_r - \mathbf{a}_r\|^2) \tag{9}$$

and let the accuracy of the exemplar module be defined as

$$EA = \exp(-.5c_e \|\mathbf{t}_e - \mathbf{a}_e\|^2). \tag{10}$$

Note that because $c_m \geq 0$, the maximal values of EA and RA are 1.0. The mean accuracy, MA, of the model can be defined as

$$MA = a_g EA + (1 - a_g)RA. \tag{11}$$

The total error, then, from Equation 8 can be expressed as $E = -\log(MA)$ and takes on nonnegative values.

Gradient descent on error yields the following learning equations. The change in weight $w_{r_k r_i}$, from Rule i to

rule-module Category Node k, is

$$\Delta w_{r_k r_i} = \lambda_r \frac{(1 - a_g)\mathrm{RA}c_r}{\mathrm{MA}} (t_{r_k} - a_{r_k}) a_{r_i}, \quad (12)$$

where λ_r is a freely estimated constant of proportionality, called the *rule-module learning rate*. The change in weight $w_{e_k e_j}$, from Exemplar j to exemplar-module Category Node k, is

$$\Delta w_{e_k e_j} = \lambda_e \frac{a_g EAc_e}{MA} (t_{e_k} - a_{e_k}) a_{e_j}, \quad (13)$$

where λ_e is a freely estimated constant of proportionality, called the *exemplar-module learning rate*. The change in attention, α_i, on Dimension i is given by

$$\Delta \alpha_i = -\lambda_\alpha \sum_{e_j} \left[\sum_{e_k} \frac{a_g EAc_e}{MA} (t_{e_k} - a_{e_k}) w_{e_k e_j} \right] a_{e_j} c |h_{e_{ji}} - d_i|. \quad (14)$$

where λ_α is a freely estimated constant of proportionality, called the *attention learning rate*. Finally, the change in weight w_{g,e_j}, from Exemplar Node j to the gating node is

$$\Delta w_{g,e_j} = \lambda_g \frac{\mathrm{EA} - \mathrm{RA}}{\mathrm{MA}} a_g (1 - a_g) \gamma_g a_{e_j}, \quad (15)$$

where λ_r is a freely estimated constant of proportionality, called the *gate-node learning rate*. Although the final output of the model, $p(K)$ (Equation 6), is a linear combination of the predictions of each module, Equations 12, 13, and 14 show that the weight adjustments depend on the discrepancy between the desired output and the category-node activation in each module separately. That is, the weight change for each module is a function of the difference between the teacher values and that module's prediction. Thus, in Equation 12 the difference $(t_{r_k} - a_{r_k})$ is used, and in Equation 13 the difference $(t_{e_k} - a_{e_k})$ is used. This causes each module to learn to produce the entire output pattern on appropriate trials rather than learning to reduce a residual from the mixed output (Jacobs et al., 1991). Despite this separation of modules, the gate differentially allocates error so that each module learns to classify those stimuli for which it is best suited.

Model Fits

ATRIUM contains 12 parameters governing its performance. In fitting this model to the data from the experiments, 8 parameters were free and 4 were fixed. Table 1 summarizes the parameters and indicates whether they were free or fixed.

Parameter estimates are based on a likelihood-ratio test statistic, G^2:

$$G^2 = 2 \sum_i f_i \ln \frac{f_i}{\hat{m}_i}, \quad (16)$$

Table 1
Summary of the Parameters in ATRIUM and the Equations in Which They Are Introduced

Parameter	Description	Equation
β_1	Rule bias for the primary dimension	1
γ_r	Rule gain	1
c	Specificity of the exemplar nodes	3
β_g	Gate bias	5
$\gamma_g = 1$	Gate gain	5
ϕ	Choice probability scaling constant	6
$c_r = 1$	Cost of the rule module	9
$c_e = 1$	Cost of the exemplar module	10
λ_r	Rule module learning rate	12
γ_e	Exemplar module learning rate	13
γ_g	Gate node learning rate	15
γ_α	Attention learning rate	14

Note. Parameters shown with values are fixed. β_1 is determined by the stimuli. The remainder of the parameters are free.

where f_i is the observed frequency of responses in Cell i and $\hat{m}_i$ is the predicted frequency in Cell i. If $f_i = 0$ for a given i, the corresponding term of the sum is also 0 (Wickens, 1989, p. 36). The model itself predicts probabilities rather than frequencies (see Equation 6). These probabilities were converted to frequencies, $\hat{m}_i$, by multiplying by the marginal frequencies for each stimulus type in each block. When the frequencies in each cell are independent, the G^2 statistic is distributed as chi-square with the degrees of freedom determined by the number of cells in the table that are allowed to vary freely. In these experiments, a given set of cells may contain repeated measures from a single participant, so independence is violated. G^2 is still a useful descriptive statistic, but it cannot be compared with chi-square for inferential statistics.

ATRIUM may be considered an extension of ALCOVE (Kruschke, 1992) because the exemplar module is an implementation of ALCOVE, and by adjusting the β_g to a sufficiently high value, it can approximate ALCOVE to an arbitrary degree of accuracy. Because ALCOVE is a subset of ATRIUM, it has fewer free parameters, viz., c, ϕ, λ_e, and λ_α (see Table 1 for a description).

Fit to Experiment 1: Extrapolation Beyond Trained Instances

The models were fit using the same trial-by-trial stimuli that participants saw. The models were fit simultaneously to both the training data and the transfer data. The training data from Experiment 1 consisted of a three-way table generated by crossing 29 blocks with 4 stimulus types and 4 response types. Trials on which participants did not respond within 6 s were not included in the table. The marginal frequencies for each stimulus type within each block were fixed in the experimental design. There are, therefore, $4 \times (4 - 1) \times 29 = 348$ degrees of freedom in the training data. The transfer data consisted of a two-way table generated by crossing 50 stimulus types by 4 response types. Once again, the marginal frequencies for each stimulus type were fixed;

Table 2
Best Fitting Parameters and G^2 Values for Participants in Experiment 1

Parameter	ALCOVE	ATRIUM
c	0.59828	1.28296
γ_α	1.96770	1.96593
λ_e	0.00855	0.32163
ϕ	6.37629	4.07742
γ_r		0.87080
λ_r		0.03375
β_g		−1.78984
γ_g		0.41313
G^2		
Training	655.40	457.84
Transfer	540.59	282.51
Total	1,195.99	740.35

hence, there are $(4 - 1) \times 50 = 150$ degrees of freedom in the transfer data. The free parameters in the model being fit each use 1 degree of freedom. Three fit values can be calculated for each model: one for the training data, $G^2(df = 348 - d, N = 25{,}013)$, one for the transfer data, $G^2(df = 150 - d, N = 3{,}071)$, and an overall value, $G^2(df = 498 - d, N = 28{,}084)$, where d is the number of free parameters in the model. The best fitting parameters and corresponding G^2 values are shown in Table 2.

ALCOVE Predictions

Fit to training data. ALCOVE was fit to the data to see whether an adequate fit could be provided by the simpler exemplar-only model rather than the hybrid rule plus exemplar model, ATRIUM. Figure 10 shows that the best fit of ALCOVE to the human learning data is reasonably good. ALCOVE learned at about the same rate as the human learners; in particular, ALCOVE shows the same few blocks of rule overgeneralization in exception classification that human learners did. The greatest discrepancy between the predictions of ALCOVE and the human data appears in the rule-learning curves (left panel of Figure 10). In roughly the first 10 blocks, ALCOVE predicts a greater proportion of correct rule responses than actually occur; in the last 10 blocks, ALCOVE underpredicts correct rule responses. This lack of fit can be explained by the low specificity (c) necessary to account for, among other things, the overgeneralization of the rule. To yield overgeneralization, presentations of the exception stimuli had to activate rule exemplars. A consequence of this was that rule stimuli also activated more than one rule-exemplar in memory. In early blocks of training when the association between exemplars and category assignments were still relatively weak, this accelerated rule learning. In these early blocks, the exceptions had only a weak association to the appropriate categories, so any interference they caused when rule stimuli were presented was minimal. Later in learning, partly because of the consistently elevated presentation frequency of the exception training instances, the association strengths between the exception exemplars and their categories were greater than the rule-exemplar association strengths. Therefore, in later blocks, rule response predictions are reduced, and inappropriate exception response predictions are elevated. In summary, the low specificity required to account for rule generalization interferes with the predicted proportion of rule responses over time. This, as will be seen, is ameliorated in ATRIUM's predictions.

Fit to transfer data. Figure 11 shows the proportion of exception responses predicted by ALCOVE. A comparison of Figure 11 with the empirical response proportions in Figure 4 shows that, although the fit is good for many stimuli, there

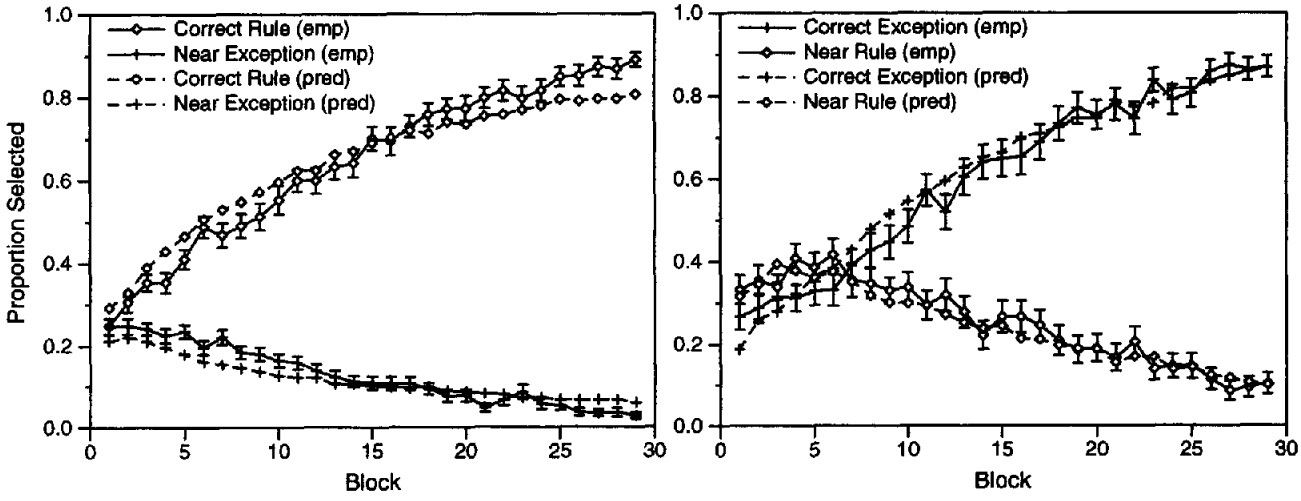

Figure 10. The left panel shows the best fit of ALCOVE to the proportion of correct rule responses and near-exception responses by block in Experiment 1. The right panel shows the best fit of ALCOVE to the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 1. In both panels, error bars extend 1 *SE* above and below the mean. emp = empirical; pred = predicted.

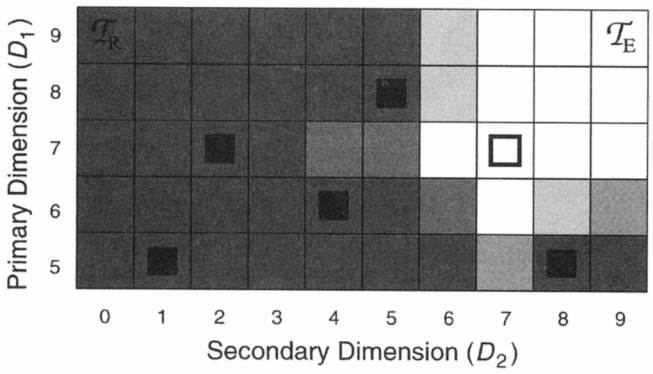

Figure 11. Proportion of exception responses in the transfer phase of Experiment 1 predicted by ALCOVE. The shading in each cell indicates the proportion of predicted exception responses. Light cells indicate a high proportion of predicted exception responses; dark cells indicate a low proportion of predicted exception responses. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript $\mathcal{T}_R$ or $\mathcal{T}_E$.

are severe discrepancies for stimuli near the exceptions. For example, the model predicts that the $\mathcal{T}_E$ stimuli (upper right cell of Figure 11) should be classified as exceptions in 43% of the trials, whereas the participants in the experiment did so in only 11% of the trials. This pattern is seen throughout the region surrounding the exceptions.

As with learning, one important factor contributing to these mispredictions is low exemplar specificity. Low exemplar specificity causes the stimuli that are only moderately similar to a training instance to activate the associated category node. Consequently, to learn the exceptions, the association weights between the exception exemplar nodes and the corresponding category nodes must be strong enough to overcome the association weights that have built up between the surrounding, partially activated exemplar nodes and the rule category nodes. Thus, during transfer, the greater magnitude of the exception association weights relative to the rule association weights and the proximity of the exception exemplar nodes to the $\mathcal{T}_E$ stimuli combine to cause the $\mathcal{T}_E$ stimuli to be classified as members of the exception category substantially more than the $\mathcal{T}_R$ stimuli.

The model does, however, predict the same attentional effects as were seen in the empirical data. This can be seen in two ways: First, exception generalization is lower for stimuli that match the exception stimulus on the secondary dimension relative to those that match the primary dimension; second, the attention weights in ALCOVE show that differences on the primary dimension are weighted more heavily than differences on the secondary one ($\alpha_1 = 1.7512$ vs. $\alpha_2 = 0.8423$). It will be seen that ATRIUM improves on the mispredictions while retaining the correct predictions.

ATRIUM Predictions

Fit to training data. Figure 12 shows the best fit of ATRIUM to the human learning data. Like ALCOVE, ATRIUM provided a good fit to the data. It can be seen that it learned at approximately the same rate as the human learners, and it shows the few blocks of overgeneralization of the rule in exception classification like human learners did. ATRIUM's predictions for rule learning are a systematic improvement over ALCOVE. The proportion of rule responses predicted by ATRIUM deviates from the empirical data substantially in only a few blocks, and only two of these are consecutive. This indicates that the deviations are less likely to be a substantial trend and more likely to be by chance. Also, the four parameters added by ATRIUM improve the learning fit

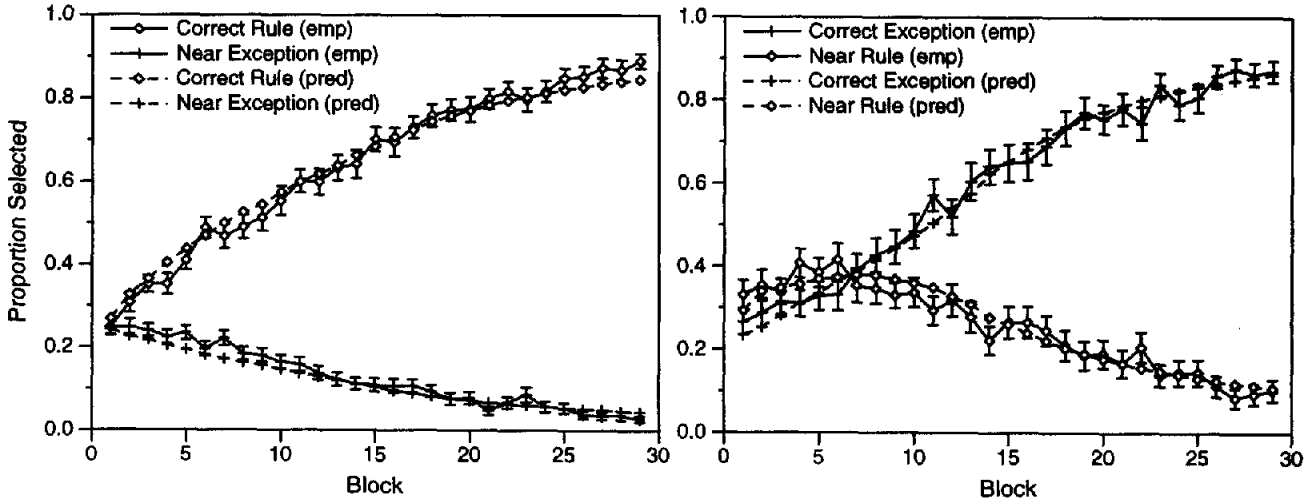

Figure 12. The left panel shows the best fit of ATRIUM to the proportion of correct rule responses and near-exception responses by block in Experiment 1. The right panel shows the best fit of ATRIUM to the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 1. In both panels, error bars extend 1 *SE* above and below the mean. emp = empirical; pred = predicted.

from $G^2(df = 344, N = 25{,}013) = 655.40$ to $G^2(df = 340, N = 25{,}013) = 457.84$. Although a statistical analysis cannot be performed because of possible violations of independence in the data, it seems unlikely that the improvement in fit is by chance.

Fit to transfer data. Figure 13 shows the proportion of exception responses predicted by ATRIUM. These predictions show a considerable improvement over the predictions of ALCOVE. In particular, ATRIUM predicts that the $\mathcal{T}_E$ stimuli should be classified as exceptions on 14% of the trials compared with ALCOVE's prediction of 43% and the empirical value of 11%. Also, just as for ALCOVE, the pattern of ATRIUM's predictions for stimuli that match the exception on either dimension shows that more attention has been allocated to the primary dimension, and the attention weights in the exemplar module show that differences on the primary dimension are more noticeable than differences on the secondary one ($\alpha_1 = 3.3361$ vs. $\alpha_2 = 1.1667$). As with learning, the fit for transfer shows improvement beyond what would be expected by chance from the addition of four parameters, $G^2(df = 146, N = 3{,}071) = 540.59$ for ALCOVE versus $G^2(df = 142, N = 3{,}071) = 282.51$ for ATRIUM.

Nevertheless, for a number of stimuli, substantial differences remain. ATRIUM overpredicts the proportion of exception responses for stimuli represented by cells horizontally and vertically adjacent to the exception. Meanwhile, it underpredicts the proportion of exception responses for stimuli that match the exception on either dimension and yet are fairly distant from the exception training stimulus. Tversky and Gati (1982) found that in certain tasks participants tend to rate two stimuli that match on a single dimension as more similar than would be predicted by models such as the exemplar subsystem in ATRIUM. One solution for this discrepancy is to modify the activation profiles of the hidden nodes within the exemplar module. Kruschke (1993a) proposed a model called APPLE whose hidden nodes cause similarity to decline abruptly for small, omni-dimensional differences between two stimuli while remaining elevated for stimuli that match on a single dimension. Another possible solution would be to use the full version of Equation 3 as presented in Kruschke (1992), so that it includes a parameter q that controls how similarity for stimuli with few differences is computed and a parameter r that controls how similarity to distant stimuli that match on dimension is computed; thus,

$$a_{e_j} = \exp\left[-.5c\sum_i \alpha_i \left|h_{e_{ji}} - d_i\right|^r\right]^{(q/r)}. \qquad (17)$$

Whereas the addition of these two parameters may improve the fit of the hidden unit activation function and participants' generalization gradient around the trained exemplars, the addition of the parameters does not affect the core issue raised by ATRIUM. That is, exemplar representation by itself is not sufficient to model human learning and extrapolation when a clear rule can be induced. This, in turn, suggests that human categorization uses rule induction, when possible, to facilitate category learning.

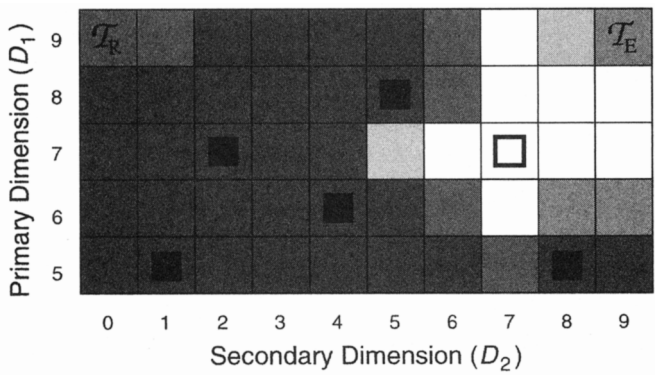

Figure 13. Proportion of exception responses in the transfer phase of Experiment 1 predicted by ATRIUM. The shading in each cell indicates the proportion of predicted exception responses. Light cells indicate a high proportion of predicted exception responses; dark cells indicate a low proportion of predicted exception responses. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript $\mathcal{T}_R$ or $\mathcal{T}_E$.

Fit to Experiment 2: Training Instance Frequency Effects

The models were fit to the data from Experiment 2 the same way they were fit to the data from Experiment 1. In Experiment 2, participants saw 16 blocks of training and 16 blocks of transfer stimuli. In this experiment, the responses to each stimulus were kept separate to retain frequency effects; hence, the data from the two frequency conditions (mixed and same frequency) were separated. This yielded a four-way table of training data that crossed the 16 blocks with the 2 experimental conditions by 20 stimulus types by 4 response types, giving $16 \times 2 \times 20 \times (4 - 1) = 1{,}920$ degrees of freedom in the training data. The transfer data were similarly organized. The key difference between the training and transfer tables was that 100 stimuli were available in transfer. This produces $16 \times 2 \times 100 \times (4 - 1) = 9{,}600$ degrees of freedom in transfer data and 11,520 degrees of freedom overall. As in Experiment 1, three fit values were obtained for each model: $G^2(df = 1{,}920 - d, N = 20{,}962)$ for training, $G^2(df = 9{,}600 - d, N = 10{,}527)$ for transfer, and $G^2(df = 11{,}520 - d, N = 31{,}489)$ for the overall fit, where d represents the number of free parameters in the model being fit. The best fitting parameters and the corresponding G^2 values are shown in Table 3.

ALCOVE Predictions

Fit to training data. The human learning data from Experiment 2 is compared with ALCOVE's predictions in Figure 14. For this experiment, ALCOVE failed to predict qualitative aspects of the data. The most obvious of these failures is its misprediction of exception training trials. When presented with exception training instances early in

Table 3
Best Fitting Parameters and G^2 Values for Participants in Experiment 2

Parameter	ALCOVE	ATRIUM
c	2.06382	12.20813
γ_α	0.02100	9.47336
λ_c	0.05718	5.32331
ϕ	3.34655	5.04795
γ_r		1.37276
λ_r		0.02600
β_g		−2.13796
γ_g		1.14772
G^2		
Training	4,839.19	3,104.93
Transfer	8,269.94	6,665.04
Total	13,109.13	9,769.97

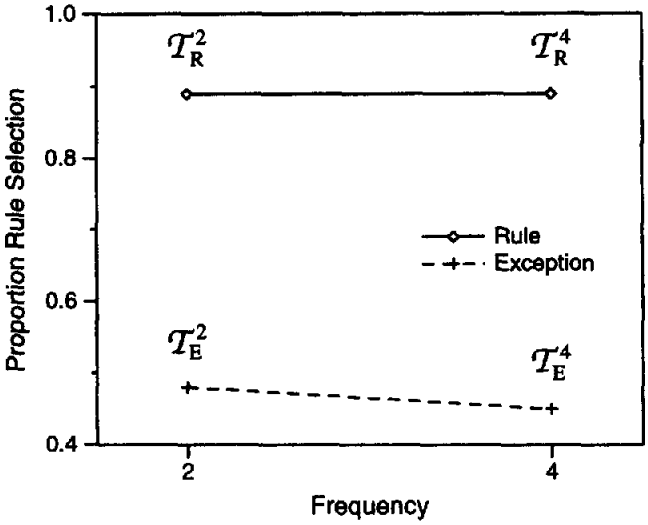

Figure 15. The proportion of appropriate rule responses predicted by ALCOVE when fit to data from Experiment 2.

training, participants tended to classify them using the rule more often than they assigned them to the correct (exception) category. Unlike the ALCOVE simulation of Experiment 1, at no point in the Experiment 2 simulation does ALCOVE predict that participants should be overgeneralizing the rule as described. Why was overgeneralization not predicted by ALCOVE with these best fitting parameter values? Part of the answer is that in Experiment 2 the generalization gradient around the exceptions is very steep. Even stimuli that were very similar to the exceptions were rarely classified as exceptions. To account for this behavior, specificity, the c parameter, must be relatively high. To show overgeneralization, however, specificity needs to be low. Hence, the two phenomena could not be accommodated by the model simultaneously.

Fit to transfer data. Comparison between the proportion of rule responses predicted by ALCOVE during the transfer trials and the results from Experiment 2 shows that, notwithstanding high specificity (c), generalization of exception responses to stimuli near the exceptions was much too strong. Moreover, ALCOVE fails to predict an effect of rule training instance frequency on generalization performance (Figure 15). It predicts that the mean proportion of rule responses for the $\mathcal{T}_R^4$ stimuli (88.5%) is slightly less than for the $\mathcal{T}_R^2$ stimuli (88.9%). The model learns to classify the high-frequency rule training instances to asymptote within the first few blocks. Thereafter, the pattern of generalization remains nearly fixed. The absence of a difference, then, may be best characterized as a ceiling effect. Whether the rule training stimulus is presented two or four times per training

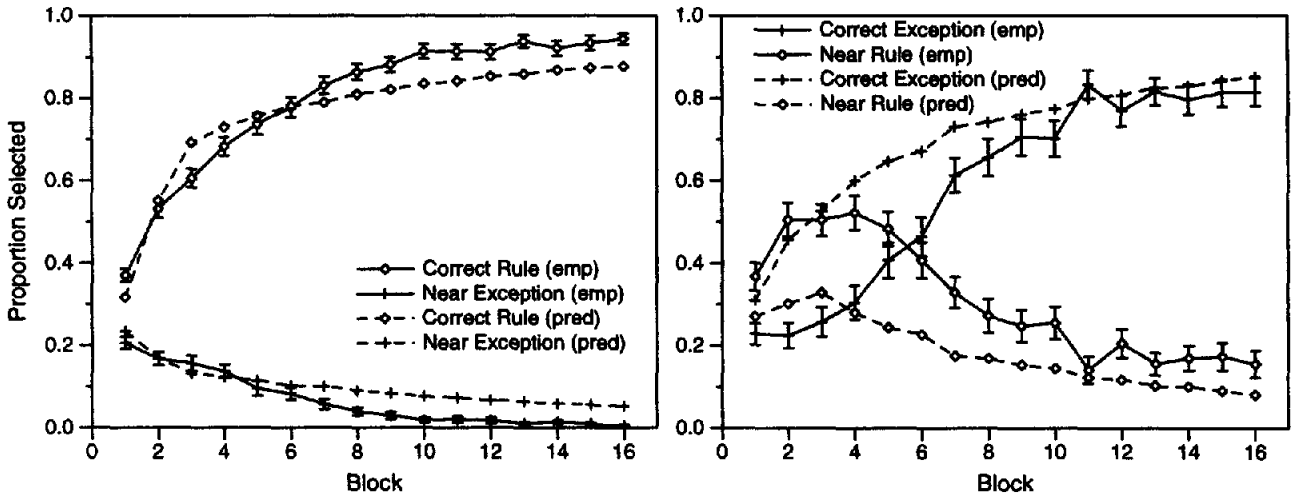

Figure 14. The left panel shows the best fit of ALCOVE to the proportion of correct rule responses and near-exception responses by block in Experiment 2. The right panel shows the best fit of ALCOVE to the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 2. In both panels, error bars extend 1 *SE* above and below the mean. emp = empirical; pred = predicted.

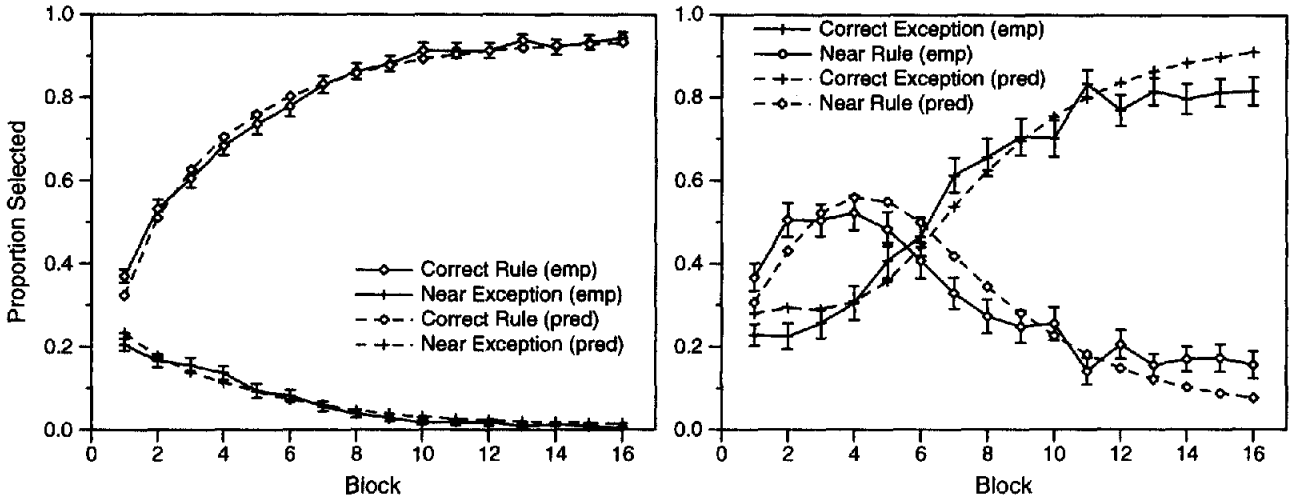

Figure 16. The left panel shows the best fit of ATRIUM to the proportion of correct rule responses and near-exception responses by block in Experiment 2. The right panel shows the best fit of ATRIUM to the proportion of correct exception responses and near-rule responses (overgeneralization) by block in Experiment 2. In both panels, error bars extend 1 *SE* above and below the mean. emp = empirical; pred = predicted.

block makes little difference to ALCOVE with the given parameter values. ALCOVE does predict a higher proportion of rule responses for $\mathcal{T}_E^2$ (48%) than for $\mathcal{T}_E^4$ stimuli (45%). Although the difference between the two predictions is in the same direction as empirical results, the magnitude of the proportions differs greatly from the human data.

In summary, ALCOVE failed to fit the data from Experiment 2. In learning and transfer, the best fit failed to capture significant qualitative aspects of the data, including overgeneralization of the rule during learning and the influence of rule training instance frequency on generalization during transfer.

ATRIUM Predictions

Fit to training data. Figure 16 shows the best fit of ATRIUM to the Experiment 2 learning data. In contrast to ALCOVE, ATRIUM provides an excellent fit to participants' responses to rule training instances and a good fit to participants' responses to exception training data. Overgeneralization of the rule when presented with exception stimuli is clearly predicted by ATRIUM. Quantitatively, however, there are discrepancies. The predicted overgeneralization appears and diminishes more slowly than the human data. Also, the model's exception categorization performance at asymptote exceeds human results.

Fit to transfer data. The predicted proportion of rule responses throughout the transfer trials in the mixed-frequency condition is shown in Appendix B. The performance of ATRIUM is far better than that of ALCOVE. Figure 17 shows the proportions of rule responses for the $\mathcal{T}$ stimuli predicted by ATRIUM. These values are all qualitatively consistent with the empirical results shown in Figure 7. ATRIUM predicts a higher proportion of rule responses for $\mathcal{T}_E^2$ stimuli (76%) relative to $\mathcal{T}_E^4$ stimuli (73%). Likewise, it predicts a higher proportion of rule responses for $\mathcal{T}_R^4$ stimuli (78%) relative to $\mathcal{T}_R^2$ (76%).

Because ATRIUM's exemplar module is freed from learning all the instances of the rule, it is able to show the effect of varying the rule training instance frequency. Because the final output of the exemplar module is mixed with the output of the rule module over the course of the experiment, the proportion of rule responses near exception training instances predicted by ATRIUM is much more consistent with the empirical data than the predictions of ALCOVE.

GENERAL DISCUSSION

Experiment 1 showed that participants' behavior is consistent with the use of a rule when they are asked to extrapolate category

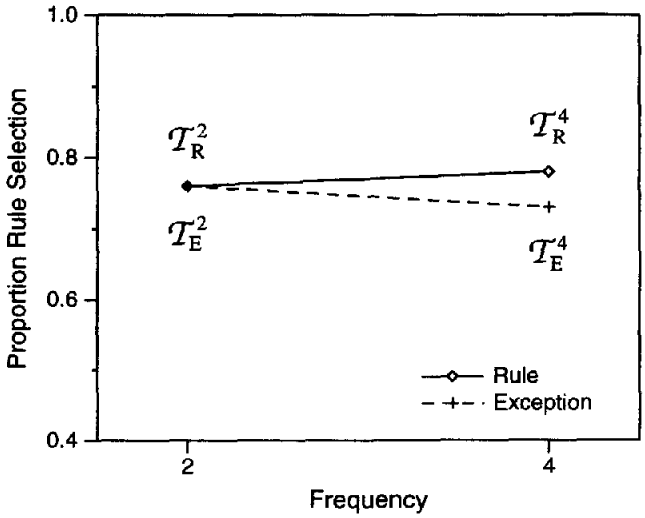

Figure 17. The proportion of appropriate rule responses predicted by ATRIUM when fit to data from Experiment 2.

knowledge beyond the trained region. Experiment 2 showed that even when it appears that participants are using a rule, properties of specific training instances influence subsequent generalization. Modeling the experimental data showed that whereas the exemplar-based model, ALCOVE, could not account for the major findings in the experiments, a hybrid rule plus exemplar model, ATRIUM, could.

We do not claim that this is the only model that can account for these data, nor do we claim that no exemplar-based model can account for these data. Our claim is that ATRIUM incorporates five principles that are in accord with the empirical results. As described previously, the empirical data suggest two representational principles: (1) rule-based representation and (2) exemplar-based representation. Moreover, the influence of each of these representations is different for different stimuli, indicating differential representational influence on an exemplar-by-exemplar basis. We refer to this selective use of different representations as (3) *representational attention.* ATRIUM also incorporates (4) error-driven learning and (5) dimensional attention (Kruschke, 1992, 1993b). The principles underlying ATRIUM presented here serve two purposes: First, they form a foundation for understanding human classification behavior; second, they facilitate comparisons with other models.

Instance-Specific Attention Weights

One question that might be posed is whether the principles of rule-based representation and representational attention are truly necessary. That is, exemplar-based representation might be sufficient if each exemplar had its own similarity gradient. With this modification of exemplar-based representation, some exemplars could generalize broadly whereas others could be quite specific. Aha and Goldstone (1990, 1992) developed an extension of the generalized context model (Nosofsky, 1984, 1986), named GCM-ISW, that does this by learning different dimensional attention weights for each exemplar and each category label. We adapted ALCOVE to incorporate this modification and fit this extended version of ALCOVE to the data from Experiment 1 using the same method as in the previous fits. Even so, the best fitting version of this model predicted that participants should classify the $\mathcal{T}_E$ stimuli as exceptions on 26% of the transfer trials compared with an empirical value of just 11%.

The improvement in this model's performance over that of ALCOVE may be best attributed to the ability of the exemplars representing training stimuli near the $\mathcal{T}_E$ stimuli to make individual attention adjustments that maximize their similarity to other members of the same category and minimize their similarity to members of other categories. In particular, by increasing the attentional weights on both dimensions around the exemplars that represented the exception training instances, this model could effectively increase the specificity of just those exemplars. Notwithstanding, the rule training exemplars near the exception made similar attentional adjustments when the exception training instances were presented (although these were to some extent neutralized by the presentation of nearby training instances from the same rule category). Thus, even though each exemplar can adjust its specificity to reduce error, the degree to which these adjustments affect the classification of the $\mathcal{T}_E$ stimuli is reduced by interference between rule and exception training instances. Although this model provided a better fit to the transfer data, without the principles of representational attention and rule-based representation, the model still could not adequately predict participants' responses to the $\mathcal{T}_E$ stimulus.

Rule Selection Mechanisms

As it is currently realized, ATRIUM does not fully implement the principle of representational attention. Although it does select between rule- and exemplar-based representations, it should also be able to select between rules on different dimensions and be able to adjust the thresholds of those rules. Two methods of implementing such a system present themselves. One method would be to extend the gating system currently used in the model. As described by Jacobs et al. (1991), the gating module can be used to select between a number of different experts. Each different dimensional rule would be implemented as a separate expert, and over time, the model would learn to choose the correct one as it adjusts the rule threshold using error-driven learning. A second method would be to expand the rule module to implement the rule selection mechanism of Busemeyer and Myung (1992). This mechanism, in turn, consists of two parts: an adaptive network that learns which rule to apply and a hill-climbing model that adjusts the parameters of the rules to maximize correct responses.

In the two experiments described here, there is no a priori reason to select either of these two rule-selection mechanisms. Considering the work of Aha and Goldstone (1990, 1992), however, one might prefer the mechanism proposed by Jacobs et al. (1991). Aha and Goldstone (1990, 1992) showed that participants have the ability to use different rules in different regions of psychological space. Using the exemplar-based gate currently implemented in ATRIUM, the mechanism of Jacobs et al. (1991) would allow each rule to be used in that region where it was best suited.

In both of the experiments presented previously, we found that in those conditions in which the height "equaled" the line segment position for both exception training instances (as denoted by the scales in the stimulus display), participants were more likely to classify all stimuli with equal values on both dimensions as members of the exception categories. ATRIUM cannot currently account for these data because these classifications involve the conjunction of two rules. Whereas one rule (e.g., "if the rectangle is taller than 4.5, classify it as a member of the 'tall' category") is already implemented, ATRIUM lacks a rule representing items that satisfy the "equal value" abstraction and a way to form conjunctions of multiple rules. If, however, an "equal-value" node and a "not-equal-value" node were added to the rule module, the rule module, acting alone, could learn to classify both the rules and the exceptions.[7] Thus, given the

[7] A possible consequence of this might be that the gate bias would shift to favor the rule module more (i.e., become more negative), thus attenuating frequency effects as in the empirical data shown in Figure A4.

proper set of rules, ATRIUM would very likely be able to select among those rules to provide an account for the data from the "equal-value" conditions.

Rule Plus Exception Model

The principle of rule-based representation is supported further by informal protocols obtained in our experiments. On completing the experiment, participants described using an if–then rule to perform the classification unless they recognized the stimulus as one of the exceptions. For example, they said that they classified all tall rectangles into one category and all short rectangles into another unless they recognized the rectangle as one of the two exceptions. Participants' protocols indicated that their decision-making processes were generally rule based unless exemplar memory overrode their rule-governed classification.

On the basis of these protocols, participants' behavior might seem to be well described by the rule plus exception model (RULEX) of category learning developed by Nosofsky, Palmeri, and McKinley (1994; Palmeri & Nosofsky, 1995). According to RULEX, people categorize by finding either dimensional rules or conjunctions of dimensional rules. If necessary, these rules are supplemented by memorized exceptions.

These mechanisms, however, might not be sufficient to account for the results of Experiment 2. In particular, the evidence for exemplar representation of rule training stimuli in Experiment 2 is beyond the scope of RULEX (as applied to categorization), because RULEX would not show training instance frequency effects. In extending RULEX to predict recognition memory performance, however, Palmeri and Nosofsky (1995) coupled it with exemplar-based representation and a parameter that weights the influence of each representational mechanism. This extension could serve as a basis for forming a RULEX plus exemplar model of categorization that might account for the results of Experiment 2.

Alternatively, ATRIUM's performance may be compared to tasks in which RULEX performs well. For example, Nosofsky et al. used RULEX to predict the distribution of participants' patterns of generalization for the category structure used by Medin and Schaffer (1978) in their Experiments 2 and 3. They compared empirical results with the predictions of RULEX and the predictions of the context model (Medin & Schaffer, 1978), and found that RULEX provided a better fit than the context model. The empirical results showed that the two most frequently used classification strategies were based on dimensional rules and the next most frequent strategy was based on similarity. RULEX showed a strong preference for the rule-based strategies, whereas the context model showed a strong preference for the similarity-based strategy. Because of its hybrid rule- and exemplar architecture, the predictions of ATRIUM might be able to address these data. Nosofsky et al., however, also analyzed the consistency of participants' classifications over three blocks of transfer trials and found that participants tended to be fairly consistent: Most participants changed two or fewer of the seven classifications between blocks. RULEX predicted similar results, whereas the context model predicted a high degree of within-participant variability. This is because the variability in RULEX is the result of stochastic selection of rules and exceptions, and the variability in the context model is the result of a probabilistic response rule. The probabilistic response rule used by ATRIUM is like that of the context model, but ATRIUM adds an additional parameter, ϕ (Equation 6), that can make the predicted probabilities more extreme and, hence, more consistent given the deterministic nature of ATRIUM. This increased within-participant consistency might come at the cost of between-participants variability, thus preventing a deterministic version of ATRIUM from adequately modeling participants' performance. One way to remedy this lack of fit might be to use a stochastic gating node as described by Jacobs et al. (1991). In the stochastic formulation of the gating node, a_g (Equations 5 and 6) does not weight the input of each module; it represents the probability that a module is chosen. Between-participants variability might, therefore, be enhanced by random processes early in training, thus allowing a higher value of ϕ to attenuate within-participant variability during transfer.

Thus, RULEX and ATRIUM address complementary aspects of classification behavior: ATRIUM addresses differentially gated mixtures of rules and exemplars but does not yet address individual differences. RULEX emphasizes the explanatory power of simple rules supplemented with the additional storage of occasional exceptions, showing that these principles can account for group as well as individual behavior. Both models are moving toward a synthesis of these behaviors.

Parallel Rule Activation and Rule Synthesis Model

Vandierendonck (1995) proposed the parallel rule activation and rule synthesis model (PRAS) of categorization, which can be profitably considered relative to RULEX and ATRIUM inasmuch as it also accounts for behavior using rule and exemplar representation. It adds to the representational framework of RULEX by including similarity gradients and continuous dimensions, and it adds to the representational framework of ATRIUM by including a mechanism for abstracting new rules. In PRAS, rules and exemplars are both represented within a homogeneous production system and are, thus, treated equivalently. In PRAS, rules are generated by connecting previously learned exemplars into a rectangular rule region in psychological space. Because rules and exemplars are treated equivalently, however, PRAS is unlikely to be able to account for the results of either of our experiments. To account for participants' pattern of generalization during transfer in Experiment 1, a model must learn to activate a rule within a broad region in the psychological space while activating an exemplar within this rule region to classify an exception correctly. PRAS could be applied to this task in two ways: First, if the probability (π, a free parameter) of generating a rule were low, PRAS would act as an exemplar model. We have shown, however, that this would lead to incorrect predictions about participants' classifications of the $\mathcal{T}$ stimuli (shown in Figure 2). Second, if π were high, PRAS could abstract a rule that could

classify the $\mathcal{T}$ stimuli correctly; however, without representational attention, the rule could not be learned well enough. Every time the exception was presented, the association between the rule and the correct rule category would be diminished, whereas the association between the rule and the exception category would be strengthened. Yet in Experiment 1, participants learned to classify the rule training stimuli faster and better than they learned to classify the exception stimuli. Moreover, if the association strength for the exception exemplar is greater than for the rule, even with rule representation PRAS would probably not be able to account for participants' pattern of generalization during transfer either. Therefore, to address the learning of exceptions, PRAS might benefit from incorporating some form of representational attention as does ATRIUM.

Explicit Rule Instructions

It is informative to consider other factors that may influence participants' performance. For example, Nosofsky et al. (1989) showed that participants' classification patterns could be modified by instructions. In their experiments, participants were presented with 16 stimuli that varied along separable dimension and were to be categorized into two rule-defined categories. Of the 16 stimuli, 7 were used as training instances and the remaining 9 were used to test generalization. In one condition, participants were instructed to classify each stimulus into one of the two categories. In the two remaining conditions, participants were given instructions to classify the stimuli according to different explicit rules. In the first condition, participants' pattern of classification was best fit by an exemplar-based model. In the latter two conditions, participants' performance was fit better by a rule- rather than an exemplar-based representation. In one of the two rule conditions, however, the rule-based model mispredicted participants' performance on instances that were highly similar to a training instance from the other category. That is, participants appeared to use exemplar-based representation even when given an explicit rule. For these data, Nosofsky et al. achieved the best fits to the data by a model that probabilistically mixed results from rule- and exemplar-based representations. Because the two representations could only be mixed uniformly throughout the stimulus space and across all trials, this model can be considered an implementation of rule- and exemplar-based representation without representational attention.

The empirical and modeling results from Nosofsky et al. (1989) suggest that rule and exemplar representation played a part in participants' classification process. By reflecting the different instructional conditions in initial parameter settings, a model based on the same architecture as ATRIUM might fit the data from all three experimental conditions. Moreover, because it incorporates representational attention, this version of ATRIUM might better predict participants' classification behavior for transfer stimuli that are highly similar to training instances.

Brooks and colleagues (Allen & Brooks, 1991; Regehr & Brooks, 1993) also showed violations of explicit verbal rules in categorization performance. In these experiments, participants were given explicit rules and practiced applying these rules to classify training stimuli. After training, participants were instructed to classify novel stimuli using the same rule. In instances in which a novel stimulus was highly similar to a training instance from the opposite category, participants tended to misclassify the stimulus. Because of the similarity between the novel stimulus and the training stimulus, participants tended to use an exemplar-based rather than a rule-based classification scheme in these cases. The principles incorporated into ATRIUM provide a basis for understanding this behavior. Although the rule module can accurately classify all the stimuli, the exemplar module still learns associations between training stimuli and category labels. As training progresses, the associations between these trained exemplars and the gate increase as well. Later, when similar novel stimuli are presented, their similarity to the training instance causes increased misclassifications by shifting attention to the exemplar module.

Competition Between Verbal and Implicit Systems Model

Ashby et al. (in press) have also explored issues surrounding verbal rules. They proposed a model named COVIS (COmpetition between Verbal and Implicit Systems) that bears some resemblance to ATRIUM. Like ATRIUM, COVIS consists of two modules that compete to produce the correct response. One module serves to categorize according to explicit verbal rules. These verbal rules, like those in ATRIUM, divide psychological space based on a single dimensional value. The other module categorizes "implicitly," using GRT (Ashby, 1988; Ashby & Gott, 1988; Ashby & Perrin, 1988; Ashby & Townsend, 1986). The two modules are gated according to their "confidence" in their responses for the given stimulus (the log-likelihood ratio of the estimated category distributions) and weight parameters for each module that are learned by a modified version of the delta rule that incorporates momentum (Rumelhart, Hinton, & Williams, 1986) and learning rate annealing (Darken & Moody, 1992).

Alfonso-Reese (1996) examined the dynamical behavior of the rule boundaries used by participants in a category learning task and compared it with that of the boundaries predicted by COVIS during learning. She found that rule boundaries of individual participants showed large, discrete jumps early in training and more incremental changes at later stages of learning. Because it uses error-driven learning in conjunction with multiple, discrete decision bounds, COVIS predicts similar behavior. In early stages of learning, COVIS selects among the rule bounds on each dimension with roughly equal likelihood. Thus, like the human learners, it exhibits large, discrete jumps early in learning. Over time, the model learns which rule bound can best account for the classification, and as error is reduced, it settles down to a fairly stable state. ATRIUM might not address these jumps in its present form.

Kruschke (1996a; Kruschke & Erickson, 1995), however, described the principle of rapid shifts of attention that, when applied to representational attention, might also account for

this behavior but in a different way. Kruschke previously applied rapid shifts of attention to different stimulus features rather than to different types of psychological representation. In category learning experiments, he found that when given feedback, participants shifted attention away from stimulus features that conflicted with previous knowledge and toward distinctive features that are consistent with previous knowledge. One essential difference between the behavior of COVIS and a model implementing rapid shifts of representational attention is that the former is a global learning adjustment, whereas representational attention is specific to individual stimuli. A possible problem with a global learning adjustment is that it is likely to cause catastrophic interference if novel stimuli are presented in later phases of learning (Kruschke, 1993a, 1993b).

It is also useful to consider whether COVIS can account for the empirical data from our Experiments 1 and 2. Because both modules in COVIS categorize according to regional boundaries, COVIS lacks exemplar-based representation. Moreover, whereas the modules in ATRIUM compete to categorize those instances for which each is best suited, the modules in COVIS compete to solve the entire categorization task individually. Thus, COVIS also lacks representational attention. Without the principle of exemplar representation, COVIS cannot account for the empirical results presented here. Exceptions cannot be classified without memory for specific instances and COVIS has no such capability as currently implemented. Furthermore, without representational attention, it is doubtful that COVIS would show the same sorts of interactions between rule and exemplar representation as participants did in our experiments.

Categorization and Language

Palermo and Howe (1970) used categorization experiments to provide an experimental analogy to learning past tense inflection. In their experiments, Palermo and Howe showed participants two-digit sequences, and participants gave one of seven single-letter responses: three regular responses and four irregular responses. For regular stimuli, participants only needed to attend to the second digit to give the correct response. Three digits mapped to each response. To recognize the four irregular stimuli, however, participants had to attend to both digits, and each irregular stimulus had its own response. Within each block of 22 trials, 12 randomly selected regular stimuli and the 10 irregular stimuli were presented. One of the irregular stimuli was shown four times per block, one was shown three times, one twice, and one once. Palermo and Howe suggested that participants' performance learning the regular and irregular stimuli from this paradigm would be analogous to learning past tense inflection for regular and irregular verbs. If this analogy between category and language learning holds, then the results of the experiments in this article should apply to language learning. For example, Experiment 1 addressed the relative influence of regular versus irregular stimuli and found that participants generalize more broadly on the basis of regular rather than irregular stimuli. Experiment 2 addressed the influence of different relative presentation frequencies and found that elevated presentation frequency causes more robust generalization for both regular and irregular stimuli.

Pinker (1991) discussed relevant linguistic data in an explanation of rulelike processes in language. For example, Pinker adduced work by Berko (1958) to show that children generally apply the regular *add -ed* inflection when given novel words like *rick* to produced *ricked,* whereas in only relatively few instances do people apply some irregular inflection (e.g., producing *splung* as the past tense of *spling;* Bybee & Moder, 1983). Thus, the results from Experiment 1 do map loosely to linguistic phenomena: People generalize more broadly from regular stimuli than from exceptions. The connection between the results from Experiment 2 and linguistic behavior, however, is more problematic. Pinker described different effects of word frequency for regular and irregular verbs. For low-frequency irregular verbs, adults rate their past tense forms (e.g., *smote, slew, bade*) as less natural than their regularized counterparts (e.g., *smited, slayed, bidded*). This is not true, however, for regular verbs. Native speakers of English find the past tense of low-frequency regular verbs no less natural than the present tense. Pinker also referred to quantitative data showing that, on the one hand, participants' perception of the naturalness of verbs' past tense form is positively correlated with the frequency of the past tense form for irregular verbs, but on the other hand, it is not for regular verbs (after partialing out the naturalness ratings for the stems).

Experiment 2 showed that the frequency of presentation can affect behavior for both regular (rule) and irregular (exception) stimuli in category learning data, whereas in linguistic data verb frequency affects adult speakers' performance for irregular verbs only. This discrepancy may be explained in at least three related ways. First, linguistic knowledge is "overlearned," whereas category knowledge in our experiments reaches only a minimal criterion. Second, the category learning tasks in our experiments are simple and can be learned in about an hour, whereas language is complex and must be learned over the course of several years. Extensive learning of a complex domain may involve different or additional processes than rapid learning in a simple domain. Third, even though the task of considering regular versus irregular forms in language and in categorization may seem comparable, the two processes may be subserved by different neural regions (see, e.g., Jaeger et al., 1996; Smith, Patalano, Jonides, & Koeppe, 1996).

Nevertheless, the same issues of representation and interaction that have been raised by models of categorization have also been addressed by models of linguistic behavior. Rumelhart and McClelland (1986), for example, proposed a homogeneous connectionist model of past tense inflection that exhibited many aspects of human past tense inflection. On the basis of the model's success, Rumelhart and McClelland challenged the idea that rules were necessary for forming the past tense. Pinker and Prince (1988), however, criticized much of their methodology and claimed that, because of their methodological problems, Rumelhart and McClelland had failed to show rules to be inessential.

Similar disputes have arisen concerning visual word

recognition, which appears to follow general phonetic rules that may be superseded by exceptions. Coltheart, Curtis, Atkins, and Haller (1993), for example, have argued that the most plausible account of visual word recognition requires one system with a set of regular transformation rules generated on the basis of exposure to a corpus of written words with their correct pronunciation and another system that memorizes the pronunciation of words that do not follow the regular transformations. In contrast, Seidenberg, Plaut, Petersen, McClelland, and McRae (1994) have proposed a number of homogeneous connectionist models that learn to generate phonemic output when presented with orthographic word representations (see also Plaut & McClelland, 1993; Seidenberg & McClelland, 1989). Seidenberg et al. claimed that these models solve the word recognition task using a single mechanism rather than using separate processes to recognize regular words or exceptions.

Might a model such as the one described by Seidenberg et al. (1994) contradict our claim that rule- and exemplar-based representations are *both* necessary to categorize rule plus exception stimuli? We believe that it does not, largely because of differences between the domain of language and the domain of category learning. Unlike the information needed to solve the categorization tasks in our Experiments 1 and 2, linguistic information cannot be accurately described in terms of a simple two-dimensional psychological space. Even limiting representation to a subset of possible orthographic elements, Seidenberg et al. use a 108-dimensional input vector. In part because of the complexity of linguistic information, then, language is learned slowly relative to the categorization tasks in our experiments. Strategies such as the application of a single one-dimensional rule that are useful when learning low-dimensional category structures might provide little help in a task as complicated as word recognition. Whereas large portions of our categorization tasks can be learned by shifting attention to a single dimension, linguistic tasks may need to be learned incrementally.

One of the shortcomings of ALCOVE that prompted the development of ATRIUM was that ALCOVE could not learn rules as fast as human participants (Kruschke & Erickson, 1994). Because of its high-dimensional architecture, a model like that used by Seidenberg et al. adapted to a category learning situation may also not be able to generalize as rapidly as humans when dimensional rules are available. Moreover, as Kruschke (1992, 1993b) showed, homogeneous, linear-sigmoid–based connectionist models can also learn to classify based on a cutoff value on a derived dimension. In particular, in Experiment 1, the exception training stimuli are linearly separable from the rule training stimuli. A common solution of the categorization problem in Experiment 1 by a homogeneous, linear-sigmoid–based connectionist model, therefore, would be to divide the exceptions from the rules along diagonal boundaries (i.e., a derived dimension) and to divide the two rule categories from each other with another boundary. Such a solution, however, predicts that in transfer trials the T_E stimuli would be classified as exceptions, whereas the results from Experiment 1 showed that participants did so on only about 11% of the trials. It is likely, then, that a model with the same homogeneous architecture as the one described by Seidenberg et al. (1994) would make these same erroneous predictions.

Conclusion

In sum, human categorization behavior is well described by a modular model that incorporates both rule and exemplar representations. The combination of rules and exceptions in categorization tasks is important for assaying these two representational systems. Nevertheless, exemplar representation is used for both rule and exception instances, so exemplar representation should not be thought of as exception representation. A key element in correctly modeling categorization in tasks such as these is capturing the interaction between the two representational structures using representational attention.

References

Aha, D. W., & Goldstone, R. (1990). Learning attribute relevance in context in instance-based learning algorithms. In M. Piattelli-Palmarini (chair), *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 141–148). Hillsdale, NJ: Erlbaum.

Aha, D. W., & Goldstone, R. (1992). Concept learning and flexible weighting. In J. K. Kruschke (Ed.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.

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

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

Ashby, F. G. (1988). Estimating the parameters of multidimensional signal detection theory from simultaneous ratings on separate stimulus components. *Perception & Psychophysics, 44,* 195–204.

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., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (in press). A neuropsychological theory of multiple systems in category learning. *Psychological Review.*

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

Ashby, F. G., & Lee, W. W. (1992). On the relationship among identification, similarity, and categorization: Reply to Nosofsky and Smith (1992). *Journal of Experimental Psychology: General, 121,* 385–393.

Ashby, F. G., & Lee, W. W. (1993). Perceptual variability as a fundamental axiom of perceptual science. In S. C. Masin (Ed.), *Advances in psychology: Vol. 99. Foundations in perceptual theory* (pp. 369–399). Amsterdam, The Netherlands: North-Holland/Elsevier.

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

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

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95,* 124–150.

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

Berko, J. (1958). The child's learning of English morphology. *Word, 14,* 150–177.

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.

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language, 59,* 251–270.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition, 21,* 413–423.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

Darken, C., & Moody, J. E. (1992). Toward faster stochastic gradient search. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems* (pp. 1009–1016). San Mateo, CA: Morgan Kaufman.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, Mark A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986.

Ervin, S. M. (1964). Imitation and structural change in children's language. In E. G. Lenneberg (Ed.), *New directions in the study of language.* Cambridge, MA: MIT Press.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178–200.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 444–458.

Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: A computational approach. *Psychonomic Bulletin and Review, 4,* 299–309.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

Jaeger, J. J., Lockwood, A. H., Kemmerer, D. L., Van Valin, R. D., Jr., Murphy, B. W., & Khalak, H. G. (1996). A positron emission tomographic study of regular and irregular verb morphology in english. *Language, 72,* 451–497.

Kalish, M., & Kruschke, J. K. (1997). Decision boundaries in one dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1362–1377.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12,* 4–34.

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

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

Kruschke, J. K. (1993b). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: The psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego, CA: Academic Press.

Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 3–26.

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science, 8,* 201–223.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Kruschke, J. K., & Erickson, M. A. (1995). *Five principles for models of category learning* [On-line]. Unpublished manuscript. Available: World Wide Web URL: http://www.indiana.edu/~kruschke/fiveprinc_abstract.html.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115–129.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

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. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

Nosofsky, R. M. (1988a). On exemplar-based exemplar representations: Repy to Ennis (1988). *Journal of Experimental Psychology: General, 117,* 412–414.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279–290.

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

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. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego, CA: Academic Press.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Palermo, D. S., & Howe, H. E., Jr. (1970). An experimental analogy to the learning of past tense inflection rules. *Journal of Verbal Learning and Verbal Behavior, 9,* 410–416.

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

Pinker, S. (1991). Rules of language. *Science, 253,* 530–535.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Plaut, D. C., & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In W. Kintsch (Ed.), *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Erlbaum.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122,* 92–114.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, England: Cambridge University Press.

Rips, L. J., Schoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1–20.

Rosch, E., & Lloyd, B. B. (Eds.). (1978). *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

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

Rosch, E. H., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 491–502.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L., & McRae, K. (1994). Nonword pronunciation and model of recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1177–1196.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–447.

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

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of "dot-pattern" classification and recognition. *Journal of Experimental Psychology: General, 121,* 278–304.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Smith, E. E., Patalano, A. L., Jonides, J., & Koeppe, R. A. (1996, November). *PET evidence for different categorization mechanisms.* Paper presented at the 37th Annual Meeting of the Psychonomic Society, Chicago, IL.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89,* 123–154.

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

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum.

Wittgenstein, L. (1953). *Philosophical investigations.* New York: Macmillan.

(Appendixes follow)

Appendix A

Results From Equal-Values Conditions

This appendix contains the results from those conditions in which the rectangle height and segment position of the exception stimuli matched (i.e., Height 2 and Position 2 or Height 7 and Position 7). These conditions were excluded from the main analysis because they caused some participants to induce a different interpretation of the stimulus structure rather than the "rule and exception" interpretation we intended. The data described in this section show evidence that some participants used a combination of the unidimensional rule and an "equal-value" abstraction to classify the exceptions.

A theoretical treatment of how people extract multiple, complex abstractions is beyond the scope of this article, although we do discuss how an extension of ATRIUM might potentially apply to these data in the General Discussion.

Experiment 1: Extrapolation Beyond Trained Instances

Training

During training, these participants' performance on rule training stimuli rose from 28% to 87% correct, whereas exception responses to these stimuli fell from 25% to 5% (see the left panel in Figure A1). Their performance on exception training stimuli rose from 27% to 86% correct, whereas rule responses to these stimuli fell from 29% to 12%. In this condition, participants categorized exception stimuli as if they were rule stimuli more than chance in Blocks 2–8.

Transfer

In this condition, participants gave a higher proportion of exception responses to T_E ($M = .33$, $SD = 0.47$) than to T_R ($M = .14$, $SD = 0.35$), $t(83) = 3.4745$, $p = 0.0008$. This follows from the earlier finding that many these participants are using an equal-value abstraction to classify exceptions. Figure A2 shows graphically the proportion of exception responses to all test stimuli; it can be seen that the positive diagonal going through the exception training instance has noticeably more exception responses (i.e., lighter shading) than in Figure 4.

Experiment 2: Training Instance Frequency Effects

Training

During training in this condition, participants' performance on rule training stimuli rose from 36% to 91% correct (left panel of Figure A3). Exception classification performance started at 24% correct and rose to 79% (right panel of Figure A3). We used two different measures to test overgeneralization. First, we compared the proportion of participants' rule responses to the proportion of their exception responses to exception training stimuli. Participants gave reliably more rule responses through Block 5 ($M = .16$, $SD = .57$), $t(54) = 2.0779$, $p = 0.04$. Participants gave reliably more rule responses than chance (.25) through Block 8 ($M = .11$, $SD = .33$), $t(54) = 2.4678$, $p = 0.02$.

Over the course of training, these participants classified stimuli that appeared four times per block better than those that appeared twice per block: 70% versus 59% correct, $F(1, 54) = 54.68$, $MSE = 0.0696$, $p < .0001$ (see Footnote 6). In this condition, performance was reliably enhanced by presentation frequency for both rule and exception training instances. The mean difference in the proportion of correct responses between the Frequency 4 and Frequency 2 rule training instances was .08 ($SD = .15$), $t(54) = 4.2024$, $p < .0001$, and the mean difference between the Frequency 4 and Frequency 2 exception training instances was .14 ($SD = .14$),

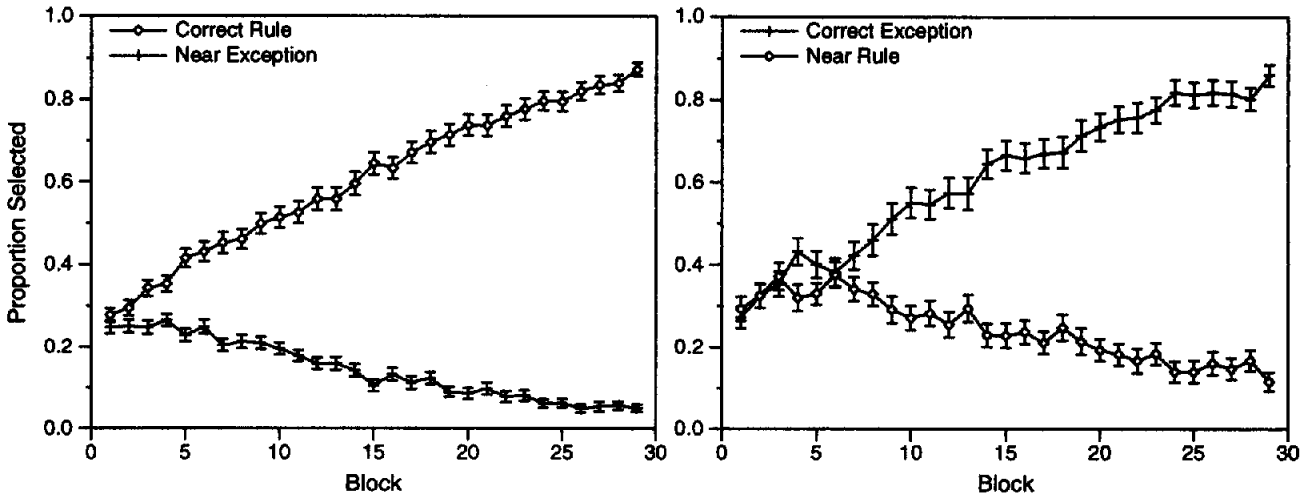

Figure A1. The left panel shows the proportion of correct rule responses and near-exception responses by block in the equal-value conditions in Experiment 1. The right panel shows the proportion of correct exception responses and near-rule responses (overgeneralization) by block in the equal-value conditions in Experiment 1. In both panels, error bars extend 1 *SE* above and below the mean.

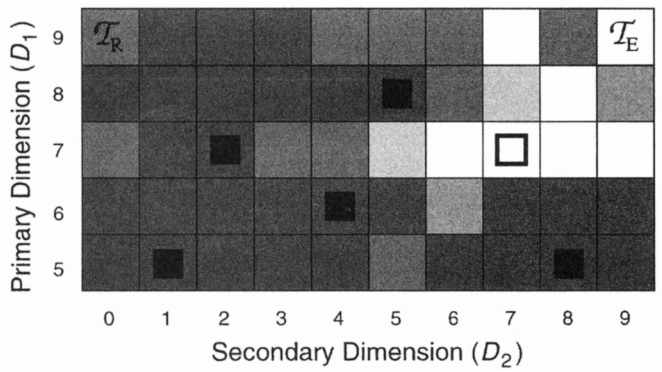

Figure A2. Proportion of exception responses in the transfer phase in the equal-value condition of Experiment 1. The shading in each cell indicates the proportion of exception responses. Light cells indicate a high proportion of exception responses; dark cells indicate a low proportion of exception responses. This diagram shows the top half of the category structure for Experiment 1. Stimuli from the bottom half have been appropriately transformed and combined with those in the top half to generate this diagram. Training instances are marked with subscript R or E (rule or exception, respectively), and the test stimuli described in the text are marked with $\mathcal{T}_R$ or $\mathcal{T}_E$.

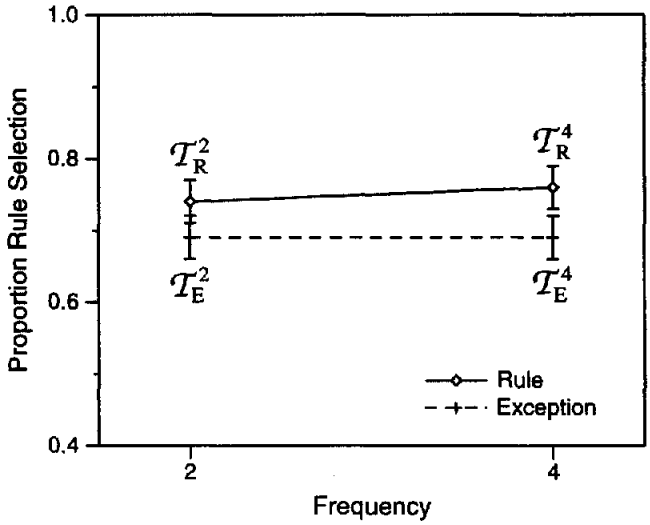

Figure A4. The proportion of appropriate rule responses for the $\mathcal{T}$ stimuli in the equal-value condition of Experiment 2.

$t(54) = 7.3982$, $p < .0001$. The facilitation for these exception instances was marginally greater than the facilitation for the Frequency 4 over the Frequency 2 rule training instances, $F(1, 54) = 3.59$, $MSE = 0.0482$, $p = .06$.

Transfer

The proportion of appropriate rule responses for the four different types of $\mathcal{T}$ stimuli is shown in Figure A4. In this condition, participants did not significantly change the proportion of rule responses in accord with variations in presentation frequency. They did not give rule responses significantly more when presented with $\mathcal{T}_R^4$ stimuli ($M = .76$, $SD = .19$) than when presented with $\mathcal{T}_R^2$ stimuli ($M = .74$, $SD = .21$). The mean of the arcsine transformed differences was 0.04 ($SD = 0.38$), $t(54) = 0.70$, $p = .48$. When presented with $\mathcal{T}_E^4$ stimuli ($M = .69$, $SD = .19$), they did not give fewer rule responses than when presented with $\mathcal{T}_E^2$ stimuli ($M = .69$, $SD = .21$). The mean of the arcsine transformed differences was 0.01 ($SD = 0.38$), $t(54) = 0.26$, $p = .79$.

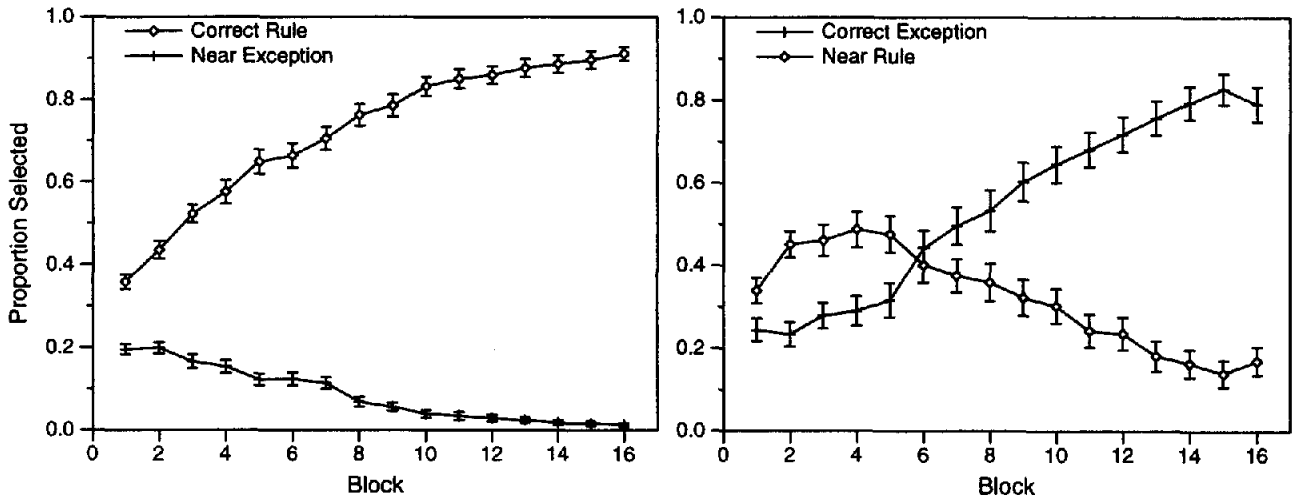

Figure A3. The left panel shows the proportion of correct rule responses and near-exception responses by block in the equal-value conditions in Experiment 2. The right panel shows the proportion of correct exception responses and near-rule responses (overgeneralization) by block in the equal-value conditions in Experiment 2. In both panels, error bars extend 1 *SE* above and below the mean.

(*Appendixes continue*)

Appendix B

Transfer Results

This appendix provides the numerical response proportions for transfer trials in both Experiments 1 and 2. Analysis of these data was provided in the main text.

Table B1 shows the proportion of exception responses in the transfer phase from the sum-to-nine conditions of Experiment 1. Table B2 shows the proportion of exception responses in the transfer phase from the equal-value conditions of Experiment 1. Table B3 shows the proportion of exception responses in the transfer phase from the sum-to-nine conditions of Experiment 1 compared with the predictions made by ALCOVE. Table B4 shows the proportion of exception responses in the transfer phase from the sum-to-nine conditions of Experiment 1 compared with the predictions made by ATRIUM.

Table B5 shows the proportion of exception responses in the transfer trials from the sum-to-nine condition of Experiment 2. Table B6 shows the proportion of exception responses in the transfer trials from the equal-value condition of Experiment 2. Table B7 shows the proportion of exception responses in the transfer trials from the sum-to-nine condition of Experiment 2 compared with the predictions made by ALCOVE. Table B8 shows the proportion of exception responses in the transfer trials from the sum-to-nine condition of Experiment 2 compared with the predictions made by ATRIUM.

Table B1
Transfer Data From the Sum-to-Nine Condition of Experiment 1

	D_2									
D_1	0	1	2	3	4	5	6	7	8	9
9	$.10_{\mathcal{T}_R}$	.06	.08	.08	.08	.08	.21	.20	.18	$.11_{\mathcal{T}_E}$
8	.15	.05	.05	.08	.05	.03■	.19	.34	.23	.08
7	.05	.13	.10■	.16	.18	.24	.32	.81□	.39	.42
6	.11	.05	.03	.02	.05■	.02	.16	.21	.10	.11
5	.03	.00■	.03	.03	.02	.05	.02	.23	.08■	.08

Note. Mean proportion of exception responses given for each stimulus. This table shows the top half of the category structure for Experiment 1. Data from the bottom half have been rotated and combined with those in the top half to generate this diagram. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript $\mathcal{T}_R$ or $\mathcal{T}_E$.

Table B2
Transfer Data From the Equal-Value Condition of Experiment 1

	D_2									
D_1	0	1	2	3	4	5	6	7	8	9
9	$.14_{\mathcal{T}_R}$	.10	.08	.12	.15	.13	.19	.35	.19	$.33_{\mathcal{T}_E}$
8	.07	.04	.07	.05	.05	.05■	.15	.27	.33	.22
7	.18	.11	.12■	.17	.17	.29	.35	.81□	.43	.38
6	.02	.04	.02	.02	.04■	.02	.23	.12	.12	.12
5	.00	.01■	.02	.01	.01	.13	.04	.06	.00■	.07

Note. Mean proportion of exception responses given for each stimulus. This table shows the top half of the category structure for Experiment 1. Data from the bottom half have been rotated and combined with those in the top half to generate this diagram. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript $\mathcal{T}_R$ or $\mathcal{T}_E$.

Table B3
Comparison Between ALCOVE*'s Predictions and Empirical Values in the Transfer Phase of Experiment 1*

D_1/ source	D_2 0	1	2	3	4	5	6	7	8	9
9										
ALC	.11	.07	.05	.07	.06	.10	.30	.55	.51	.43
Emp	$.10_{T_R}$	.06	.08	.08	.08	.08	.21	.20	.18	$.11_{T_E}$
8										
ALC	.07	.03	.02	.03	.03	.06	.30	.66	.56	.48
Emp	.15	.05	.05	.08	.05	.03■	.19	.34	.23	.08
7										
ALC	.12	.07	.08	.10	.14	.21	.53	.87	.73	.64
Emp	.05	.13	.10■	.16	.18	.24	.32	.81□	.39	.42
6										
ALC	.05	.02	.01	.02	.01	.03	.18	.45	.29	.25
Emp	.11	.05	.03	.02	.05■	.02	.16	.21	.10	.11
5										
ALC	.05	.02	.01	.02	.01	.03	.10	.26	.10	.13
Emp	.03	.00■	.03	.03	.02	.05	.02	.23	.08■	.08

Note. ALC = ALCOVE; Emp = empirical. Mean proportion of exception responses given for each stimulus. This table shows the top half of the category structure for Experiment 1. Data from the bottom half have been rotated and combined with those in the top half to generate this diagram. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript T_R or T_E.

Table B4
Comparison Between ATRIUM*'s Predictions and Empirical Values in the Transfer Phase of Experiment 1*

D_1/ source	D_2 0	1	2	3	4	5	6	7	8	9
9										
ATR	.08	.08	.07	.07	.06	.07	.11	.24	.19	.14
Emp	$.10_{T_R}$	.06	.08	.08	.08	.08	.21	.20	.18	$.11_{T_E}$
8										
ATR	.07	.06	.05	.05	.04	.04	.12	.49	.33	.24
Emp	.15	.05	.05	.08	.05	.03■	.19	.34	.23	.08
7										
ATR	.07	.05	.04	.06	.07	.20	.50	.86	.60	.49
Emp	.05	.13	.10■	.16	.18	.24	.32	.81□	.39	.42
6										
ATR	.07	.05	.05	.05	.04	.05	.10	.28	.14	.15
Emp	.11	.05	.03	.02	.05■	.02	.16	.21	.10	.11
5										
ATR	.07	.04	.05	.05	.05	.05	.06	.13	.05	.07
Emp	.03	.00■	.03	.03	.02	.05	.02	.23	.08■	.08

Note. ATR = ATRIUM; Emp = empirical. Mean proportion of exception responses given for each stimulus. This table shows the top half of the category structure for Experiment 1. Data from the bottom half have been rotated and combined with those in the top half to generate this diagram. Training instances are marked with a filled or open square (rule or exception, respectively), and the test stimuli are marked with a subscript T_R or T_E.

(Appendixes continue)

Table B5
Transfer Data From the Sum-to-Nine Condition of Experiment 2

Rectangle height	Segment position									
	0	1	2	3	4	5	6	7	8	9
9	.77	.80	$.80_{R^1_■}$	.83	.77	.87	.80	$.89_{R^1_■}$	.86	.89
8	.82	.78	.65	.76	.88	$.85_{R^1_■}$	.84	.79	.84	.88
7	$.81_{R^1_■}$	.80	$.22_{E^4_□}$	.68	.87	.80	.86	$.86_{R^2_■}$	.85	$.83_{R^1_■}$
6	.72	.73	.63	.78	$.78_{R^1_■}$	.75	.75	.78	.83	.84
5	.62	$.60_{R^1_■}$	.57	.75	.61	.80	$.74_{R^1_■}$	.80	.80	.69
4	.76	.71	.78	$.79_{R^1_●}$	.80	.74	.69	.59	$.65_{R^1_●}$	.72
3	.89	.93	.91	.84	.89	$.78_{R^1_●}$	.82	.60	.83	.77
2	$.95_{R^1_●}$	.96	$.87_{R^4_●}$	.95	.87	.91	.77	$.27_{R^2_○}$	.73	$.86_{R^1_●}$
1	.93	.92	.90	.94	$.92_{R^1_●}$	.87	.87	.67	.91	.90
0	.94	.89	$.91_{R^1_●}$	.91	.81	.90	.82	$.84_{R^1_●}$	.88	.91

Note. Mean proportion of exception responses given for each stimulus. Training instances are marked with an R or an E (rule or exception, respectively), a shape to represent the correct category response, and a superscript indicating the relative frequency.

Table B6
Transfer Data From the Equal-Value Condition of Experiment 2

Rectangle height	Segment position									
	0	1	2	3	4	5	6	7	8	9
9	.77	.79	$.70_{R^1_■}$	.81	.72	.78	.80	$.79_{R^1_■}$	.80	.81
8	.80	.75	.72	.67	.77	$.70_{R^1_■}$	.77	.82	.76	.78
7	$.70_{R^1_■}$	.75	$.29_{E^4_□}$	.68	.77	.70	.80	$.84_{R^2_■}$	.76	$.81_{R^1_■}$
6	.70	.67	.67	.68	$.68_{R^1_■}$	.79	.79	.83	.73	.75
5	.67	$.63_{R^1_■}$	.57	.62	.58	.65	$.67_{R^1_■}$	.68	.71	.76
4	.65	.69	.63	$.64_{R^1_■}$	.71	.58	.66	.62	$.70_{R^1_●}$	.66
3	.75	.84	.63	.70	.75	$.78_{R^1_●}$	.62	.67	.77	.68
2	$.77_{R^1_●}$	.78	$.84_{R^4_●}$	.77	.76	.78	.77	$.41_{E^2_○}$	.80	$.80_{R^1_●}$
1	.82	.84	.86	.80	$.85_{R^1_●}$	.86	.80	.81	.72	.86
0	.80	.84	$.89_{R^1_●}$	.86	.82	.85	.81	$.84_{R^1_●}$	.84	.73

Note. Mean proportion of exception responses given for each stimulus. Training instances are marked with an R or an E (rule or exception, respectively), a shape to represent the correct category response, and a superscript indicating the relative frequency.

Table B7
Comparison Between ALCOVE*'s Predictions and Empirical Values in the Transfer Trials in Experiment 2*

Rectangle height/ source	Segment position									
	0	1	2	3	4	5	6	7	8	9
9										
ALC	.66	.65	.54	.61	.79	.87	.87	.88	.84	.77
Emp	.77	.80	.80$_{R^1_■}$	.83	.77	.87	.80	.89$_{R^1_■}$	.86	.89
8										
ALC	.68	.55	.28	.48	.80	.90	.92	.91	.88	.82
Emp	.82	.78	.65	.76	.88	.85$_{R^1_■}$	.84	.79	.84	.88
7										
ALC	.65	.38	.12	.32	.75	.87	.93	.91	.91	.79
Emp	.81$_{R^1_■}$	.80	.22$_{E^4_□}$	.68	.87	.80	.86	.86$_{R^2_■}$	.85	.83$_{R^1_■}$
6										
ALC	.69	.57	.24	.45	.77	.86	.90	.86	.78	.69
Emp	.72	.73	.63	.78	.78$_{R^1_■}$	.75	.75	.78	.83	.84
5										
ALC	.66	.65	.36	.36	.59	.72	.80	.68	.48	.40
Emp	.62	.60$_{R^1_■}$	.57	.75	.61	.80	.74$_{R^1_■}$	.80	.80	.69
4										
ALC	.38	.45	.70	.79	.75	.65	.44	.48	.70	.69
Emp	.76	.71	.78	.79$_{R^1_●}$	.80	.74	.69	.59	.65$_{R^1_●}$	.72
3										
ALC	.60	.73	.86	.87	.84	.78	.57	.45	.68	.75
Emp	.89	.93	.91	.84	.89	.78$_{R^1_●}$	.82	.60	.83	.77
2										
ALC	.81	.91	.94	.91	.85	.77	.46	.24	.49	.65
Emp	.95$_{R^1_●}$	.96	.87$_{R^4_●}$	.95	.87	.91	.77	.27$_{E^2_○}$	.73	.86$_{R^1_●}$
1										
ALC	.78	.90	.94	.92	.87	.78	.59	.40	.63	.68
Emp	.93	.92	.90	.94	.92$_{R^1_●}$	.87	.87	.67	.91	.90
0										
ALC	.71	.81	.87	.86	.80	.75	.65	.61	.67	.70
Emp	.94	.89	.91$_{R^1_●}$	.91	.81	.90	.82	.84$_{R^1_●}$	.88	.91

Note. ALC = ALCOVE; Emp = empirical. Mean proportion of exception responses given for each stimulus. Training instances are marked with an R or an E (rule or exception, respectively), a shape to represent the correct category response, and a superscript indicating the relative frequency.

(*Appendixes continue*)

Table B8
Comparison Between ATRIUM's Predictions and Empirical Values in the Transfer Trials in Experiment 2

Rectangle height/ source	Segment position									
	0	1	2	3	4	5	6	7	8	9
9										
ATR	.77	.79	.86	.77	.79	.83	.77	.86	.78	.81
Emp	.77	.80	$.80_{R^1_\blacksquare}$	.83	.77	.87	.80	$.89_{R^1_\blacksquare}$	.86	.89
8										
ATR	.79	.78	.69	.74	.78	.86	.79	.80	.78	.81
Emp	.82	.78	.65	.76	.88	$.85_{R^1_\blacksquare}$	.84	.79	.84	.88
7										
ATR	.84	.74	.24	.75	.79	.76	.75	.83	.79	.80
Emp	$.81_{R^1_\blacksquare}$	.80	$.22_{E^4_\square}$	.68	.87	.80	.86	$.86_{R^2_\blacksquare}$	.85	$.83_{R^1_\blacksquare}$
6										
ATR	.72	.71	.64	.70	.79	.72	.73	.71	.70	.72
Emp	.72	.73	.63	.78	$.78_{R^1_\blacksquare}$	.75	.75	.78	.83	.84
5										
ATR	.52	.62	.53	.51	.53	.53	.64	.53	.53	.54
Emp	.62	$.60_{R^1_\blacksquare}$	.57	.75	.61	.80	$.74_{R^1_\blacksquare}$	.80	.80	.69
4										
ATR	.58	.56	.58	.66	.60	.62	.58	.58	.68	.59
Emp	.76	.71	.78	$.79_{R^1_\bullet}$	.80	.74	.69	.59	$.65_{R^1_\bullet}$	.72
3										
ATR	.74	.73	.75	.75	.73	.82	.73	.69	.76	.77
Emp	.89	.93	.91	.84	.89	$.78_{R^1_\bullet}$	.82	.60	.83	.77
2										
ATR	.89	.82	.89	.80	.81	.84	.80	.44	.79	.89
Emp	$.95_{R^1_\bullet}$	.96	$.87_{R^4_\bullet}$	.95	.87	.91	.77	$.27_{E^2_\bigcirc}$	.73	$.86_{R^1_\bullet}$
1										
ATR	.82	.82	.86	.82	.88	.80	.81	.76	.84	.84
Emp	.93	.92	.90	.94	$.92_{R^1_\bullet}$	.87	.87	.67	.91	.90
0										
ATR	.83	.81	.89	.82	.80	.81	.80	.90	.81	.84
Emp	.94	.89	$.91_{R^1_\bullet}$	.91	.81	.90	.82	$.84_{R^1_\bullet}$	.88	.91

Note. ATR = ATRIUM; Emp = empirical. Mean proportion of exception responses given for each stimulus. Training instances are marked with an R or an E (rule or exception, respectively), a shape to represent the correct category response, and a superscript indicating the relative frequency.

Appendix C

Scaling Study

To fit ATRIUM to the data from Experiments 1 and 2, psychological coordinates of the stimuli were derived in a separate scaling study.

Procedure

Participants read instructions for the experiment on a computer screen. They were told that their task was to rate the similarities of rectangles on a scale ranging from 1 to 9. During the instructions, they were shown two pairs of rectangles and told that each pair of rectangles was of average similarity and should be rated 5. The distance between each member of each pair in the physical stimulus space was 3, and each pair varied along only one dimension. Participants were encouraged to use the whole range of the scale as they made their judgments. They were also encouraged to make each judgment carefully and were given time to rest between each trial.

The two stimuli on each trial were presented sequentially for 1.5 s each. After a 100-ms delay, a screen with a scale from 1 to 9 was shown to prompt participants to rate the similarity of the two rectangles. The scale was labeled with 1 (*least similar*) and 9 (*most similar*).

We assumed that the scaling solution would be rectangular. This reduced the number of trials necessary to constrain the solution. Also, pilot studies indicated that ratings on extremely dissimilar rectangles showed greater variability than other ratings, so the maximum metric distance between stimulus pairs in this experiment was limited to 5. This also reduced the number of trials. Participants saw two different types of trials. In one type, the two stimuli varied along only one dimension. These constrained the relative distances within each dimension. In the second type, the two stimuli varied along both dimensions (see Figure C1). The primary purpose of these was to determine the relative salience of each dimension.

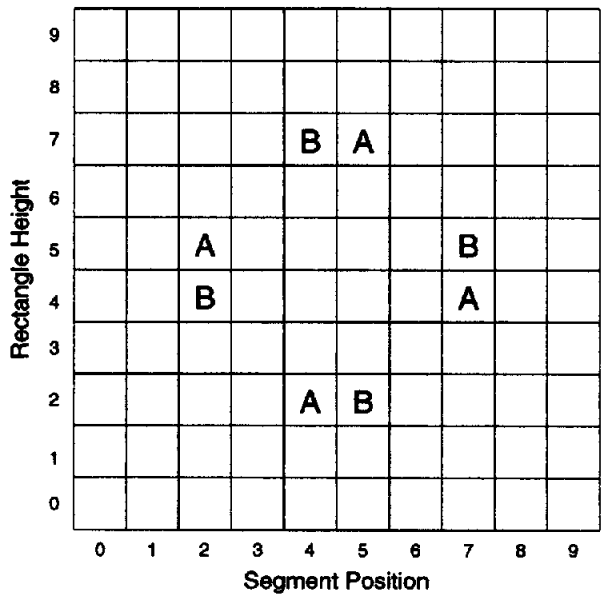

Figure C1. The layout of the stimuli used for two-dimensional comparisons in the scaling study. The stimuli labeled A were compared with one another, and the stimuli labeled B were compared with one another.

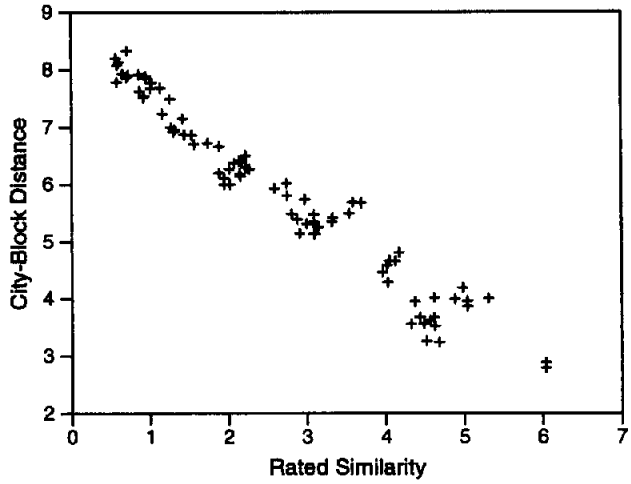

Figure C2. Correspondence between similarity ratings and computed psychological distance. Distance accounts for more than 95% of the variance in similarity ratings.

Each participant made 35 × 2 dimensions × 2 orders = 140 one-dimensional judgments and 6 × 2 sets × 2 repetitions × 2 orders = 48 two-dimensional judgments. The stimuli used for the one-dimensional judgments were chosen randomly, and the order in which the stimulus pairs were displayed was randomized for each participant.

A total of 36 participants took part in the study for partial credit in an introductory course at Indiana University Bloomington.

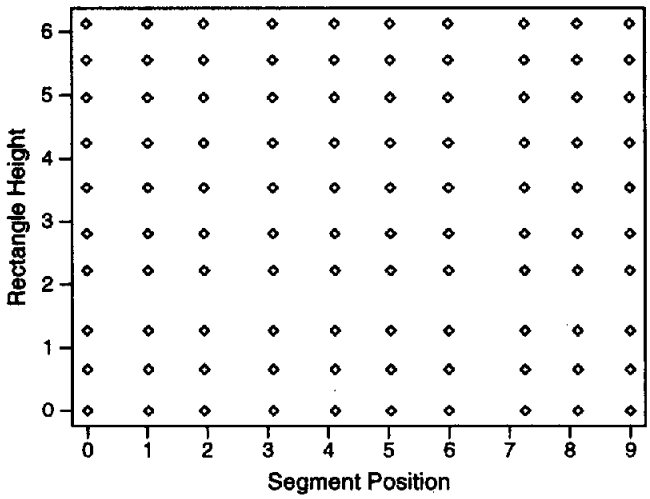

Figure C3. Scaling study results.

(*Appendixes continue*)

Results and Scaling Solution

The similarity ratings used to generate the psychological positions of the stimuli were computed by averaging across participants, presentation order, and repetitions. Each one-dimensional similarity was, therefore, the mean of 72 judgments, and each two-dimensional similarity was the mean of 144 judgments.

Following Kruskal (1964), psychological coordinates were chosen to minimize stress between the city block distance between pairs and a prediction of proximity as a monotone function of similarity ratings. Stress was computed as $\Sigma_i(d_i - \hat{d}_i)^2 / \Sigma_i d_i^2$, where d_i is the city block distance between the current psychological coordinates of Pair i and $\hat{d}_i$ is the monotonic proximity prediction derived from the mean similarity ratings for Pair i. Under the assumption that the psychological dimensions were independent, each psychological dimension had 10 values, just as each physical dimension had 10 values. Because the location in physical space and the overall scale were arbitrary, the psychological values corresponding to the physical value of 0 were fixed at 0.0, and the psychological value of Segment Position 9 was fixed at 9.0. Thus, there were 17 free parameters to fit the distances in psychological space to the similarity ratings.

The correlation between the rated similarities and the computed distances is shown in Figure C2. The best fit had a stress of 0.061. Distance accounted for 95.17% of the variance in the similarity ratings. The best fitting psychological coordinates of the stimuli are shown in Figure C3, where it can be seen, for example, that height has a smaller psychological range than horizontal segment position.

Received November 27, 1996
Revision received April 29, 1997
Accepted May 12, 1997 ■

AMERICAN PSYCHOLOGICAL ASSOCIATION
SUBSCRIPTION CLAIMS INFORMATION

Today's Date:_____

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do **NOT** duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL)

ADDRESS

DATE YOUR ORDER WAS MAILED (OR PHONED)

CITY STATE/COUNTRY ZIP

____PREPAID ____CHECK ____CHARGE
CHECK/CARD CLEARED DATE:_____

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

YOUR NAME AND PHONE NUMBER

ISSUES: ___MISSING ___DAMAGED

TITLE	VOLUME OR YEAR	NUMBER OR MONTH

Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4–6 weeks.

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: _____ **DATE OF ACTION:** _____
ACTION TAKEN: _____ **INV. NO. & DATE:** _____
STAFF NAME: _____ **LABEL NO. & DATE:** _____

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.