

Rule-based Extrapolation in Perceptual Categorization

Michael A. Erickson
Carnegie Mellon University

John K. Kruschke
Indiana University, Bloomington

Erickson and Kruschke (1998) provided a demonstration that in certain situations people will classify novel stimuli according to an extrapolated rule, even when the most similar training exemplar is an exception to the rule. This result challenged exemplar models. Nosofsky and Johansen (2000) have called this finding into question by offering an exemplar-based explanation for those data based on the perceptual features of the stimuli. Here, we describe the results of a new experiment that yields results similar to those found previously without the questionable perceptual features: Participants who learn to classify all the training stimuli have patterns of generalization that indicate a combination of rule and exemplar representation. ATRIUM, a hybrid rule and exemplar model (Erickson & Kruschke, 1998), is shown to account for these data much better than ALCOVE, an exemplar model (Kruschke, 1992). Moreover, four alternate exemplar explanations, including one suggested by Nosofsky and Johansen, cannot account for our new findings.

Perceptual categorization is central to cognition. To generalize from one situation to the next, observers must be able to recognize the essential similarities between different stimuli. As a simple example, although the letters “a” and “ɑ” have different perceptual properties, people treat the two letters identically for the purpose of understanding the meaning of a written word. Because of the fundamental nature of this behavior, psychologists have spent a great deal of effort trying to understand how people learn to categorize items in their environment.

Many psychological and folk theories hold that *rules* play a substantial role in people’s categorization behavior (e.g., Brooks, Squire-Graydon, & Wood, 1998; Bruner, Goodnow, & Austin, 1956). This belief probably has several sources. Among these is that most people have had the experience of sorting items according to instructions (e.g., sorting laundry by color). To the degree that people are successful at grouping items according to rule-derived instructions, they show evidence of rule use in categorization (Allen & Brooks, 1991; Nosofsky, Clark, & Shin, 1989). Further, many people are familiar with puzzles in which the assigned task is to guess the rule (e.g., is “N” the next character in the sequence

“OTTFSSSEN”?). Based upon their success in these tasks, people may conclude that they induce rules in more general classification behavior.

In contrast to theories positing rule use in categorization, other theories hold that people’s classification behavior can be explained by a system that generalizes from previously seen *exemplars* (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). According to these theories, people make categorization decisions in two steps: In the first step, the similarity between the stimulus in question and exemplars of previously experienced stimuli stored in memory is determined. Then in the second step, the stimulus is assigned to the category to which it is most similar overall. Exemplar models have provided highly successful accounts of people’s behavior in inductive category-learning tasks (e.g., Choi, McDaniel, & Busemeyer, 1993; Kalish & Kruschke, 1997; Kruschke, 1992, 1993a, 1993b, 1996; Medin & Schaffer, 1978; Nosofsky, 1986, 1987, 1988, 1989; Nosofsky et al., 1989; Nosofsky, Gluck, Palmeri, McKinley, & Gauthier, 1994; Nosofsky & Kruschke, 1992; Nosofsky, Kruschke, & McKinley, 1992; Nosofsky & Palmeri, 1996; Palmeri, 1999).

In fact, exemplar theories have been so successful that some theorists have proposed that they are sufficient to account for people’s behavior across a range of categorization tasks. For example, Nosofsky and Johansen (2000) stated that “in free-strategy situations in which people learn categories via induction over training exemplars... a single-system, exemplar-similarity approach appears adequate to account for the major phenomena of interest” (p. 395). Before addressing this claim further, it is important not to overestimate its scope. It refers specifically to those situations in which people learn to classify stimuli into categories from trial-by-trial experience and in which they are given no instructions that summarize the nature of the category structure (cf. Allen & Brooks, 1991; Grings, Schell, & Carey, 1973; Lewandowsky, Kalish, & Griffiths, 2000; Noelle & Cottrell, 1996, 2000; Nosofsky et al., 1989). As will be explained

Michael A. Erickson, Department of Psychology, Carnegie Mellon University and Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania, and John K. Kruschke, Department of Psychology and Cognitive Science Program, Indiana University, Bloomington.

This work was supported in part by National Institute of Mental Health (NIMH) Research Training Grant PHS-T32-MH19983, NIMH National Research Service Award 1-F32-MH12722, and NIMH FIRST Award 1-R29-MH51572.

Correspondence concerning this article should be addressed to Michael A. Erickson, Center for the Neural Basis of Cognition, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213. Electronic mail may be sent to erickson@cnbc.cmu.edu.

in more detail later, in these tasks, the participant's job is to learn a mapping between the stimuli and a set of labels. On each trial, an individual stimulus is presented, and the participant is instructed to assign the correct category label. Once the participant makes a response, feedback is given indicating the correct label.

In this article we argue that in these free-strategy category learning situations, when possible, people do induce rules to accelerate learning. We argue that although exemplar representation is a very general and useful learning strategy, human learners recognize and take advantage of situations in which broad generalizations are possible in a way that cannot be accounted for by exemplar theories alone. Although we have previously found support for this proposition in perceptual category learning tasks (Erickson, 1999; Erickson & Kruschke, 1998; Kruschke & Erickson, 1994), Nosofsky and Johansen (2000) have argued that these results were an artifact of details in the stimulus displays (cf. Thomas, 1998). In this article, we demonstrate that the stimulus details cited by Nosofsky and Johansen are not necessary to find evidence of rule induction.

Erickson and Kruschke (1998, Experiment 1) previously provided evidence of rule induction in free-strategy category learning tasks by examining people's patterns of generalization after training had been completed. In that experiment, participants were presented with rectangular stimuli that varied along two separable dimensions: the height of the rectangle and the horizontal position of an internal line segment. Most of the stimuli were members of two categories that could be distinguished by a simple rule on one of the dimensions. The remaining training stimuli were each members of their own categories, and they could be distinguished from the members of the two rule categories by considering the values of both dimensions.

On each trial during training, a stimulus was presented, and participants were instructed to indicate the correct category label. At first, participants just guessed, but after each response, they were told the correct label. Once training was complete, Erickson and Kruschke (1998) tested participants' generalization by presenting them with stimuli not seen during training. During this phase of the experiment, no feedback was provided. In particular, Erickson and Kruschke were interested in how participants generalized to stimuli that were similar to the exceptions. They found that even for novel stimuli that were more similar to the exception training instances than to the rule training instances, participants tended to make rule-consistent classifications. Thus, participants appeared to have knowledge of a rule that could be distinguished from exemplar-based knowledge of the exceptions.

Although the behavioral data indicated that participants had induced knowledge of a rule, Erickson and Kruschke (1998) sought to corroborate this by fitting two different category-learning models to the data. The first, *ALCOVE* (Kruschke, 1992), utilized exemplar representation but had no rule representation. The second, *ATRIUM* (Erickson & Kruschke, 1998), combined the exemplar representation used in *ALCOVE* with rule representation. Using rule representation,

ATRIUM was able to provide a better overall fit to the data and, more importantly, it was able to provide a better account of participants' generalization to stimuli that were similar to the exceptions. Thus, Erickson and Kruschke concluded that participants were using a combination of rule and exemplar representation to perform this category learning task.

A demonstration that rule-and-exemplar representation provides a better account for the results of Erickson and Kruschke's (1998) Experiment 1 than an exemplar-only representation does not eliminate the possibility of alternate explanations. For example, in this experiment, the stimuli were presented with tic marks and numeric labels along the left and bottom edges of the screen so that each stimulus could be uniquely identified by two, ordered digits (e.g., "height 7, segment position 2"). Participants might have augmented their representations of the stimuli with these numeric identifications. To test this hypothesis, Nosofsky and Johansen (2000) performed a variation of this experiment in which the tic marks and numeric labels were removed from the stimuli: the rectangle with its internal line segment was presented on an otherwise empty screen. When they did so, they found that the participants' pattern of generalization no longer provided strong evidence of rule use by participants.

Although Nosofsky and Johansen (2000) were able to show that the omission of tic marks and numeric labels in this experiment changed the degree to which rule induction occurred, they did not show that the tic marks were necessary for rule induction. In other words, they did not show that the evidence of rule induction found by Erickson and Kruschke (1998) was an artifact of these stimulus details; they only showed that in one experiment, the omission of that information reduced the degree to which rule induction could be demonstrated. The experiment presented in the present article provides evidence of rule induction using stimuli that do not contain tic marks or labels identifying the values on the stimulus dimensions. Thus, the new data provide a continuing challenge to exemplar-only theories of categorization.

This article is organized as follows. We first present an experiment in which participants show a pattern of generalization that we argue is inconsistent with the predictions of extant exemplar models of categorization. We then provide explicit tests of the ability of *ATRIUM*, a hybrid rule-and-exemplar model, and *ALCOVE*, an exemplar model, to show that, indeed, the addition of rule representation provides a qualitative improvement. We conclude by discussing the implications of these findings for exemplar theorists and considering what might cause learners to induce rules.

Experiment: Rule-Based Extrapolation Beyond Exceptions

The stimuli were fixed-width rectangles whose heights could vary among eight, equally-spaced values. Each rectangle was presented with a vertical, interior line segment whose horizontal position could vary among eight equally spaced values. This yielded $8 \times 8 = 64$ possible stimuli. No tic marks or value labels were shown.

The category layout for these stimuli is shown in the left

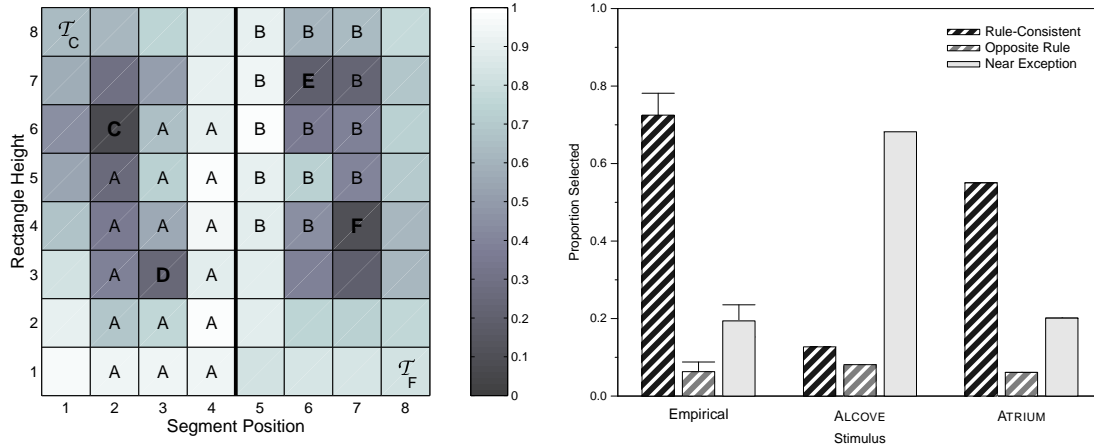


Figure 1. Category structure and results from the experiment for learners. The left panel shows the category structure and the proportion of rule-consistent Category A and B responses from the last block of the experiment. The rows and columns respectively represent the stimulus values for the rectangle heights and segment positions. The 33 cells containing a letter A–F were training instances for their respective categories. The empty cells and the cells containing \mathcal{T}_C and \mathcal{T}_F indicate the 31 transfer stimuli used to test generalization. The numbers 1–8 along the axes are used to label the different stimulus values for the purposes of this figure only and are not indicated in the stimulus presentation. Light cells indicate a high proportion of rule-consistent Category A and B responses. The right panel shows the proportion of learners’ rule-consistent, opposite-rule, and nearby exception responses for the \mathcal{T} stimuli. The first set of bars shows the empirical results, followed by ALCOVE’s and ATRIUM’s predictions. Error bars extend 1 SE above the mean.

panel of Figure 1. The members of Categories A and B can be distinguished by considering only the position of the line segment. Given that a stimulus is a member of Category A or Category B, if the segment is in the left half of the rectangle, the stimulus must be a member of Category A, and if the segment is in the right half of the rectangle, the stimulus must be a member of Category B. Thus, most of the stimuli could be distinguished by a simple one-dimensional rule.

There are four exceptions to this rule. Each exception is a member of its own category, and hence, the exceptions constitute Categories C–F. The left panel of Figure 1 shows the placement of these exceptions in this category structure. Notice that two of the exceptions, C and F, lie at the extreme edges of the regions with trained rule instances. The remaining untrained stimuli, the *transfer* stimuli, were available to test generalization. These included the stimuli labeled \mathcal{T}_C and \mathcal{T}_F . The \mathcal{T} stimuli are important because they provide a strong test of the three types of representation under consideration: rule representation, exemplar representation, and an augmented exemplar representation suggested by Nosofsky and Johansen (2000, p. 386), which posits that the value along each dimension for each stimulus is stored in memory along with a standard exemplar trace. This yields a four-dimensional stimulus space with two dimensions for the standard exemplar trace and two additional dimensions for the dimensional values. If participants are extrapolating their category knowledge using rules, they will tend to classify the \mathcal{T}_C and \mathcal{T}_F stimuli as members of Categories A and B, respectively. If they are generalizing using exemplar knowledge, they will tend to classify the \mathcal{T}_C and \mathcal{T}_F stimuli as members of Exception Categories C and F because of the high degree of similarity between the \mathcal{T} stimuli and these ex-

ceptions. Finally, if participants are using an augmented exemplar representation such as the one suggested by Nosofsky and Johansen, they will tend to classify the \mathcal{T} stimuli more often as members of the *opposite* rule category. As will be explained in more detail later, this is because (a) under this theory, the standard exemplar representation of each stimulus is augmented with its value on each dimension, and (b) each \mathcal{T} stimulus perfectly matches three members of the opposite rule category. In particular, the \mathcal{T}_C stimulus matches the augmented height value of the three Category B training instances with height 8, and the \mathcal{T}_F stimulus matches the augmented height value of the three Category A training instances with height 1. It is important to acknowledge that according to the rationale given by Nosofsky and Johansen (2000), the augmented exemplar model would not be suited for this experiment inasmuch as the stimuli are presented without numeric labels or tic marks. Nevertheless, this category structure provides the ability to detect whether or not participants are *covertly* generating such a representation. In sum, examination of participants’ generalization to the \mathcal{T} stimuli provide a qualitative assay for each of the three representations under consideration.

Method

Participants

Eighty-one Indiana University undergraduate students volunteered to participate for partial credit in an introductory psychology class.

Apparatus

The stimuli were presented on PC-compatible computers in individual, sound-dampened, dimly lit booths. Participants made responses by pressing keys on the standard keyboard.

Procedure

Participants were presented with 381 consecutive trials. The trial order was separately randomized within blocks for each participant. The first 112 trials were composed of 14 blocks of 8 trials in which each of the four exception training items (Categories C, D, E, and F) were presented twice.¹ The next 138 trials were composed of two blocks of 69 trials. In each of these blocks, the four exception training items were presented 10 times each, and the 29 rule-consistent training items (Categories A and B) were presented once each. In the last block of the experiment, 131 trials, each exception training item was presented 10 times, each rule-consistent training item was presented one time, and each transfer item was presented two times.

Participants were given written instructions on the computer screen indicating that their task would be to learn the correct label for each figure presented to them. They were also told that near the end of the experiment, they would see figures for which no label would be provided. For these figures, participants were instructed to make their best educated guess.

On each trial, a stimulus was presented and participants were instructed to respond using one of the keys D, F, G, H, J, or K. These keys were randomly mapped to categories A–F for each participant. Participants were given 30 s to respond. Once they responded, the correct category label was presented along with feedback indicating whether their response was correct, except on a transfer trial, in which case the feedback indicated merely that their response was recorded. They were then given up to 30 s to study the correct answer (if presented) and were told to press the space bar to continue.

Results

Before proceeding with other analyses, participants' performance was evaluated twice: first, to determine whether they learned to classify the exceptions correctly before the Category A and B training items were introduced, and second, to determine whether they had learned to classify the members of all six categories in the final block of the experiment. The null hypothesis used to determine chance responding for these comparisons was that participants had learned to respond by probability matching. The alpha level used was .05. For example, chance responding for Category A in the final block of the experiment would be determined by finding the critical value using a binomial distribution with $N = 16$ and $p = 16/69 \approx .23$ or $\text{bin}_{\text{crit}(\alpha=.05)}(N = 16, p = .23) = 8$. The critical values for the remaining categories in both evaluations were computed in the same way. In the first evaluation, we computed participants' performance classifying

each of the four exception training items correctly in their last 10 presentations before the Category A and B training items were introduced. The criterion used for each category was $6/10 = .6$. Four of the participants failed to classify all of the exceptions at above chance levels. In the second evaluation, we computed the remaining 77 participants' performance classifying the training items in all six categories during the final block. The criteria used were $8/16 = .5$ for Category A, $6/13 = .46$ for Category B, and $4/10 = .4$ for Categories C–F. Of the 77 participants, 40 were able to classify the members of all six categories better than chance and 37 were not.

The 40 participants who were able to classify the members of all six categories better than chance in the final block of the experiment are referred to as *learners*. The 37 participants who learned to classify all four exceptions better than chance but were unable to assimilate the Category A and B training items are referred to as *partial learners*. Inasmuch as the key comparisons for distinguishing between rule representation, exemplar representation, and the augmented exemplar representation depend upon participants knowing how to classify the training stimuli correctly, the data from the learners is most relevant for distinguishing between the theories. The partial learners' results, therefore, are presented only briefly and are not considered further.

Empirical Results

Partial Learners. As would be expected from the criteria used to designate the partial learners, during the last block, the participants responded correctly on 71% of the trials ($SD = 18.9\%$) in which exception training items were presented, but only 41% of the trials ($SD = 13.1\%$) in which rule-consistent training items were presented.

The pattern of generalization to the \mathcal{T} stimuli was considerably different for the partial-learners than for the learners, as will be seen. The modal response for the partial learners was the nearby exception category ($M = 39\%$, $SD = 34.1\%$) although they made a substantial number of rule-consistent ($M = 24\%$, $SD = 31.5\%$) and opposite-rule responses as well ($M = 27\%$, $SD = 32.5\%$; cf. the right panel of Figure 1).

Learners. The proportion of rule-consistent Category A and B classifications for the learners during the last block of the experiment is shown in the left panel of Figure 1, where light squares indicate rule-consistent Category A and B responses. During the last block, learners responded correctly on 69% of the trials in which rule-consistent training items were presented ($SD = 9.9\%$) and on 82% of the trials in which exception training items were presented ($SD = 9.6\%$).

¹ Pilot studies were performed in which the number of initial exception-only blocks were reduced. Although the performance of participants in the pilot studies who learned to classify the training stimuli correctly was not qualitatively different from the participants who learned the training stimuli in the present study, the presentation of these additional exception-only blocks increased the proportion of participants who learned the training stimuli.

Evidence of rule representation is most clearly found in participants' generalization to novel stimuli, especially the \mathcal{T} stimuli. Rule-consistent responses to these stimuli suggest rule representation whereas exception responses suggest exemplar representation, and opposite-rule responses suggest the augmented exemplar representation described previously. The right panel of Figure 1 shows these proportions of participants' responses.² The learners clearly tended to classify the \mathcal{T} stimuli according to the rule. Thus, these results provide initial evidence for rule representation in this strictly perceptual category learning experiment.

In light of this rule-consistent classification of the \mathcal{T} stimuli, a valid concern might be that this pattern was driven by learners who had mastered the rule at the expense of correct classification of the exceptions. Learners' performance on rule-consistent training items, however, is positively correlated with performance on exception training items presented during the last block of the experiment, $r = .58$, $t(38) = 4.364$, $p < .0001$. More importantly, there are no outliers that would indicate that some learners were trading off between accuracy on rule-consistent and exception training items. Nevertheless, it could still be the case that rule-consistent classification of the \mathcal{T} stimuli is most prevalent among learners who were poorest at classifying the Category C and F training stimuli. Reliable evidence of a trade-off between rule-consistent classification of the \mathcal{T} stimuli and correct classification of the Category C and F training stimuli in the last block could not be found, $r = -.07$, one-tailed $t(38) = -0.435$, $p = .33$. Hence, it is reasonable to conclude that the learners' performance was sufficiently homogeneous to be aggregated and examined jointly.

Model-Based Analyses

Although the learners' pattern of responses is suggestive of rule-based responding, it is difficult to predict how well an exemplar model of categorization can account for a pattern of results without fitting it to the data. To test whether the addition of rule representation as instantiated in ATRIUM is useful to account for participants' behavior, both an exemplar model, (ALCOVE; Kruschke, 1992), and the rule-and-exemplar model (ATRIUM; Erickson & Kruschke, 1998) were fit to these data.

ATRIUM is a hybrid model composed of modules. Each stimulus presented to the model is processed simultaneously by all of the modules. The number of modules is determined by the number of dimensions in the category structure being learned. There is one module per dimension that learns rules and one additional module that learns exemplar associations. Thus, to simulate the data from the present experiment, three modules were required. The first two were a *height-rule* module and a *position-rule* module that could each learn to classify the stimuli based on single-dimension rules. The third was an *exemplar* module that could learn to classify each stimulus by evaluating its similarity to the members of each of the categories. This exemplar module is identical to the ALCOVE model (Kruschke, 1992). The modules in ATRIUM are controlled by a gating mechanism, which also

uses exemplar representation. It learns which module is best suited to classify each stimulus and controls the degree to which each module contributes to each classification so that the the best module contributes the most. The gating mechanism also controls how rapidly each module learns based on feedback so that the module that performs the best on each classification also receives the most feedback (Jacobs, Jordan, & Barto, 1991; Jacobs, Jordan, Nowlan, & Hinton, 1991; Jacobs, 1997).

Both models were presented with the exact sequence of stimuli seen by each participant, and on each trial, they predicted the probability of choosing each of the six possible responses. ALCOVE makes its responses by considering the similarity of each stimulus to the previously presented stimuli in each of the categories. For example, if the stimulus with segment position 2 and height 1 were presented after all the training stimuli had been presented at least once, it would be highly similar to many of the Category A stimuli (especially previous instances of itself that it would match perfectly), less similar to the Category D training stimulus, and still less similar to the Category B, C, E, and F stimuli. Hence ALCOVE would predict that this stimulus would be classified most often as a member of Category A. This theory of categorization has successfully accounted for a large number of empirical phenomena (Choi et al., 1993; Kalish & Kruschke, 1997; Kruschke, 1992, 1993a, 1993b, 1996; Nosofsky et al., 1994; Nosofsky & Kruschke, 1992; Nosofsky et al., 1992; Nosofsky & Palmeri, 1996; Palmeri, 1999).

ATRIUM posits that in addition to exemplar representation, people may form dimensional rules. In this experiment, for example, ATRIUM would be expected to form a rule that classifies all stimuli whose segment is positioned to the left in Category A and all stimuli whose segment is positioned to the right in Category B. Additionally, because the gating mechanism in ATRIUM learns which stimuli are best classified by each rule and which are best classified using exemplar similarity, it can learn that the Category A and B stimuli should be classified using the segment-position rule and that the Category C, D, E, and F stimuli should be classified using exemplar similarity. This rule-and-exemplar theory of categorization has proved superior to ALCOVE in a number of instances (Erickson, 1999; Erickson & Kruschke, 1998; Kruschke & Erickson, 1994).

ATRIUM is a generalization of ALCOVE. It is more complex and, therefore, should be expected to provide a quantitatively superior account of participants' performance relative to ALCOVE. Because of this, it is important to measure the quality of each model's fit in a manner that takes ATRIUM's additional complexity into account. Akaike's information criterion statistic (AIC, Akaike, 1974) penalizes models for complexity as measured by the number of free parameters.

² The \mathcal{T}_C and \mathcal{T}_F data are collapsed because, although learners were more likely to give an exception response to the \mathcal{T}_C stimuli ($M = .35$, $SD = .455$) than to the \mathcal{T}_F stimuli ($M = .04$, $SD = .175$), $t(39) = 4.41$, $p < .001$, all three explanations under consideration make this same prediction. The distinction between \mathcal{T}_C and \mathcal{T}_F , therefore, does not help distinguish between the theories, and hence they have been combined.

Models with more free parameters incur a greater penalty so that, ideally, simpler models are preferred. Unfortunately, the degree to which this penalty is sufficient depends upon factors external to the models. In particular, as the number of data points increases, the impact of the penalty decreases so that more complex models tend to be favored inappropriately (Busemeyer & Wang, 2000; McDonald & Marsh, 1990; Myung & Pitt, 1997).

Because of the possible bias induced using the AIC, it is beneficial to use complementary methods of evaluating the models to corroborate or refute its initial indications. In this article, this is done using two different methods. The first method is simply to supplement the AIC with other fit statistics.

The second method is to examine the models' qualitative predictions for a set of *critical* data. In this case, the critical data were the number of rule, exception, and opposite-rule responses given to the two \mathcal{T} stimuli. These critical data points are important because, as described, each of three theories under consideration predicts a different pattern of responses. Although these data points were included in the set to which the models were fit, the parameter adjusting procedure has no way of "knowing" which of the data points are the critical ones. Thus, as long as the proportion of critical data points is small, they are unlikely to have a disproportionate impact on the estimated parameter values. In this case, the critical set comprised just 6 data points out of 155 (about 4%). This method of examining qualitative predictions for a set of critical data is, in many ways, more informative than quantitative measures of fit. By examining these critical data, one can determine in a broad sense the degree to which a model can or cannot account for gross aspects of behavior.

Erickson and Kruschke (1998, see also Kruschke & Erickson, 1994) have already shown that it is difficult for ALCOVE to account for some of the details of learning in experiments such as the one presented here. Therefore, if ALCOVE were fit to the learning data from this experiment, its parameter settings would reflect an attempt to simultaneously account for both the learning and the transfer data. Consequently, the quality of the fit to the transfer data would almost certainly be diminished. Hence, to maximize the possibility that ALCOVE would be able to provide a good qualitative account of the transfer data, the models were only fit to participants' patterns of classification for the 31 transfer stimuli in the last block of the experiment (shown in the left panel of Figure 1). Each of these stimuli could have been classified into one of six possible categories yielding $31 \text{ stimuli} \times (6 - 1)$ independent category choices = 155 degrees of freedom in the data. As stated previously, three of the response categories for the \mathcal{T}_C and \mathcal{T}_F stimuli were identified as the critical data. The corresponding response categories (i.e., rule, exception, and opposite rule) for the \mathcal{T} stimuli were then collapsed, and the resulting empirical data and model predictions are shown in the right panel of Figure 1.

The two models shared five common parameters. These were the learning rate for exemplar classification, λ_e ; the learning rate for dimensional attention, λ_α ; a measure of overall discriminability between the stimuli, c ; a measure of

decision certainty, ϕ ; and a measure of the relative salience of the two dimensions of stimulus variation, ν . This last parameter was necessary because although a previous similarity scaling study had shown that the two dimensions of stimulus variation used in this study were separable (Erickson & Kruschke, 1998, Appendix C), a scaling study was not performed with these particular 64 stimuli to determine the relative salience of the two dimensions. By including ν as a free parameter, the best value for each fit could be determined, and no fit would be prejudiced by a poor estimate. Consistent with the scaling study reported by Erickson and Kruschke (1998), in all the fits reported, the salience of the segment position was greater than that of rectangle height.

ATRIUM's learning was governed by five other parameters: the learning rate for rule-based classification, λ_r ; the degree of rule precision, γ_r ; the learning rate for the gating mechanism, λ_g ; the relative cost for rules based on rectangle height, c_h ; and the relative cost of rules based on segment position, c_p . For a complete description, see Erickson (1999). Both models were fit to the data from the learners and the partial-learners separately, and the parameter spaces were searched using a simulated annealing algorithm (Corana, Marchesi, Martini, & Ridella, 1987; Goffe, Ferrier, & Rogers, 1994).

ALCOVE's best fit yielded $AIC = 2624.46$, $R^2 = .465$, and $RMSD = 0.202$ (for parameter values of $\lambda_e = 0.309$, $\lambda_\alpha = 0.001$, $c = 0.736$, $\phi = 0.734$, and $\nu = 2.283$). In contrast, ATRIUM's best fit yielded $AIC = 1454.79$, $R^2 = .840$, and $RMSD = 0.110$ (for parameter values of $\lambda_e = 0.026$, $\lambda_\alpha = 0.148$, $c = 4.800$, $\phi = 11.693$, $\nu = 1.425$, $\lambda_r = 0.796$, $\gamma_r = 8.594$, $\lambda_g = 1.493$, $c_h = 7.890$, and $c_p = 1.046$). Under the AIC and RMSD, lower values indicate a better fit, whereas under R^2 , higher values indicate a better fit. Therefore, using the fit statistics alone, it can be seen that the improvement in fit provided by the additional principles instantiated in ATRIUM beyond those instantiated in ALCOVE was substantial. Moreover, examination of the models' fit to the critical \mathcal{T}_C and \mathcal{T}_F stimuli, shown in the right panel of Figure 1, indicates a fundamental problem with the predictions made by ALCOVE. ALCOVE predicts that participants should tend to classify the \mathcal{T} stimuli as members of the nearby exception categories whereas ATRIUM, in accord with the empirical data, predicts that they should tend to classify the \mathcal{T} stimuli as members of the surrounding rule categories. Therefore, both methods of evaluating the models' fits indicate that exemplar representation alone is insufficient to account for the learners' behavior.

Discussion

This combination of empirical data and modeling indicates that participants exhibited strong rule-like extrapolation of categorization. They classified novel stimuli according to a rule, even when the most similar training exemplars were exceptions to the rule. These results were accounted for by ATRIUM (Erickson & Kruschke, 1998), which is a rule-and-exemplar model, but they could not be fit by ALCOVE (Kruschke, 1992), which is an exemplar model.

Alternate Exemplar Models

It is important to acknowledge that exemplar models are quite powerful, and it may be possible that a model in that class could provide an alternate explanation for these data. Without an exhaustive search, this result would be impossible to rule out. Nevertheless, we briefly consider four possible classes of modifications that could be made to exemplar models and describe why it seems unlikely that they could account for these data.

The first class of modifications we consider are those suggested by Nosofsky and Johansen (2000) that augment exemplar representation with value-on-dimension representation. Because of the design of the category structure, relative to ALCOVE, this type of model predicts an increase in the proportion of participants' *opposite* rule category responses when presented with the \mathcal{T} stimuli. ALCOVE, however, deviates from the data because it predicts too few *rule-consistent* responses, not too few *opposite-rule* responses. Thus, this class of augmented models would not account for participants' behavior.

The second class of exemplar models we consider are those that have alternate monotonic transformations of the psychological stimulus space. The version of ALCOVE that was fit to the data from the present study allowed for proportional stretching of the stimulus space via the ν parameter and attention learning. These mechanisms allow each dimension to be stretched or shrunk uniformly. It is possible, however, that a version of ALCOVE could account for participants' patterns of generalization if it assumed that the participants' psychological representation of the stimulus space was transformed in a more complex manner. We tested this possibility by allowing the initial positions of the representation of each rectangle height and segment position in the stimulus space to be a free parameter, given the constraint that their order not change (i.e., monotonicity), while simultaneously fitting ALCOVE to the data. This version of ALCOVE was able to provide an overall fit that was comparable to ATRIUM (without freely scaled heights and positions). Whereas the overall fit as measured by the $AIC = 1402.19$ was slightly superior to that of ATRIUM, the other two measures of fit $R^2 = .757$, and $RMSD = .136$ indicated an advantage for ATRIUM (for parameter values of $\lambda_e = 0.258$, $\lambda_\alpha = 0.001$, $c = 0.197$, and $\phi = 1.583$).³ More importantly, this model failed to account for the participants' pattern of generalization of the critical \mathcal{T} stimuli. With this more flexible psychological space, this version of ALCOVE was able to increase its predicted proportion of rule-consistent responses and decrease its opposite-rule responses (.34 and .07, respectively). However, whereas the learners' modal response was *rule-consistent*, ALCOVE's was still the *near exception* category (.54; cf. the right panel of Figure 1). Thus, ALCOVE still tends to classify the \mathcal{T} stimuli as exceptions even under the best fitting monotonic transformation of the psychological space. Although the fit statistics are somewhat equivocal, examination of the critical data indicates a clear, qualitative advantage for rule representation in this case.

The third class of exemplar models we consider are those

that add biases to the outputs. Inasmuch as the different categories occur with different frequencies, it is reasonable to suppose that participants might be sensitive to these frequencies and incorporate them into their response strategy. If participants learned, for example, that Category A responses occurred 16 times per block in the last several blocks of the experiment whereas Category C responses occurred only 10 times per block, that might explain why they tend to classify the \mathcal{T}_C stimulus as a member of Category A. To test this hypothesis, we extended ALCOVE to incorporate response bias learning. The best fit of this extended version of ALCOVE yielded $AIC = 2225.66$, $R^2 = .620$, and $RMSD = .170$ (for parameter values of $\lambda_e = 0.782$, $\lambda_\alpha = 0.001$, $c = 1.021$, $\phi = 0.685$, and $\nu = 1.532$). Although this extension of ALCOVE was able to improve upon the fit of standard ALCOVE, it was still far worse than that of ATRIUM. Moreover, even though this version of ALCOVE was able to predict far fewer exception responses than the standard ALCOVE (.30), it did so by augmenting *both* rule-consistent and opposite-rule responses (.27 and .31, respectively). Thus, under both methods of model evaluation, this version of ALCOVE failed to meet the standards set by ATRIUM.

The last class of exemplar models we consider are ones that might parse the classification task into two constituent parts: a rule-consistent part and an exception part. For example, participants might have learned to recognize the exceptions and to classify them using one exemplar representation while using another exemplar representation for all other stimuli. This would permit both accurate exception classification and rule-consistent generalization. As with other possible exemplar models, we cannot rule out the possibility that some such model could account for participants' pattern of generalization in the present experiment. We postulate, however, that the key to such an account would be the coordination of the two representations. We have termed the principle in ATRIUM that underlies its coordination of different representations *representational attention*. The idea behind this principle is that people learn to use different representations in different situations. In ATRIUM, it is implemented in the exemplar-based gating mechanism that learns which representation is best suited to classify each stimulus and directs corrective feedback to this same representation. This allows rapid specialization of each of the representations and reduces interference between them. It seems likely that any model that is able to account for these data using multiple representations, even multiple exemplar representations, will need to instantiate some form of representational attention. Whether our hypothesis holds, however, remains an open question for exemplar theorists to explore.

³ In this fit, rectangle height 1 was assigned to location 0.0, and the best fitting locations of the subsequent heights were 0.289, 0.907, 4.872, 12.340, 19.151, 21.592, and 22.312. Without loss of generality, the first and last segment positions were assigned to locations 0.0 and 7.0, and the best-fitting locations of the intervening segment positions were 0.993, 1.720, 2.384, 6.174, 6.269, and 7.000.

When are rules induced?

The main focus of this article has been to demonstrate that, when possible, people use rule representation even in free-strategy inductive category learning situations. Why might learners do this? It seems likely that a principle benefit of rule induction is a dramatic acceleration of learning. Kruschke and Erickson (1994) provided evidence for this hypothesis by examining separately the performance of participants who learned to classify exception training stimuli before the rule-consistent training stimuli (*exception-first* participants) and the performance of participants who learned to classify the rule-consistent training stimuli before the exception training stimuli (*rule-first* participants). Whereas the rate of learning for the exception-first participants could be well accounted for by ALCOVE, it was unable to account for the learning performance of the rule-first participants using exemplar representation alone. It was simply unable to learn to classify the rule-consistent stimuli as fast as the human learners. An extended version of ALCOVE that incorporated many of the principles of ATRIUM, however, was able to provide an excellent account of the speed with which the rule-first participants learned to classify the rule-consistent training stimuli.

If rule-induction accelerates classification learning, why is it that evidence of rule-induction is not found in all category learning tasks in which a rule exists? In particular, why did Nosofsky and Johansen (2000) not find evidence of rule representation in their follow-up of Erickson and Kruschke's (1998) Experiment 1? According to ATRIUM, different representations (e.g., rules and exemplars) compete with one another to classify each stimulus. Learning to use rules, therefore, reduces the capacity to learn associations between exemplars and category labels. Given this capacity constraint, it makes sense that as tasks become more difficult, participants will be less likely to induce rules. It should be noted that the difficulty of the task depends both upon the task itself as well as the resources learners bring to it. Thus, it should be expected that for most tasks there will be individual differences in rule induction.

According to the prediction made by ATRIUM that rule induction should occur less as task difficulty increases, a principle reason that Nosofsky and Johansen (2000) failed to find strong evidence of rule induction should be that their task was more difficult. It seems likely that removing the tickmarks and the numeric labels from the stimulus presentations in their variation of Erickson and Kruschke's (1998) Experiment 1 did make the learning task more difficult. By reducing the amount of identifying information for each stimulus, they increased the confusability of the stimuli. This, then, would have made the formation of clear associations between stimuli and category labels more difficult. Further, there is evidence in the data that suggests that the Nosofsky and Johansen study may have been more difficult. Despite offering a \$25 premium for the top three participants, which was designed to motivate participants to perform better, Nosofsky and Johansen ultimately excluded the data from 39% of their participants. Erickson and Kruschke, on the other

hand, offered no monetary reward in their study and, using Nosofsky and Johansen's criteria for excluding data, would only have excluded 33% of their participants due to learning performance. Unfortunately, this same analysis cannot be applied to the present study due to differences in its design. For example, participants in the present study were given six possible category labels whereas in the previous studies they were given only four. Clearly, therefore, the same criteria cannot be applied. Nevertheless, many aspects of the design of the present experiment may have made it easier than Nosofsky and Johansen's study. Whereas in their study, 100 different stimuli were presented using the range of rectangle heights and line segment positions available on the computer screen, in the present study, there were only 64. This quite likely served to decrease confusability between stimuli thereby making the formation of associations between stimuli and category labels easier. Further, the training stimuli were more densely distributed across stimulus space in the present experiment. This may have served to delineate the rule more clearly. Although these explanations are only suggestive, they do provide a plausible account of the different results obtained by Erickson and Kruschke (1998), by Nosofsky and Johansen (2000), and in the present study. That is, when a category learning task becomes too difficult, people are less likely to be able to afford the cost of learning the rule. This causes them to rely on exemplar-based category learning.

Summary

ATRIUM (Erickson & Kruschke, 1998; Erickson, 1999; Kruschke & Erickson, 1994) was able to provide a significantly better account of the data from the present experiment than ALCOVE (Kruschke, 1992) and several other possible exemplar models. The combination of the empirical data and modeling results presented herein provide a continuing challenge to theorists who argue that a purely exemplar-based account of category learning is sufficient.

We argue that ATRIUM is able to account for the results from these category learning experiments because it incorporates both rule and exemplar representation. As we have shown previously (Erickson & Kruschke, 1998, Experiment 2), exemplars may be used to represent both rule and exception training instances. Therefore exemplar representation should not be thought of as *exception* representation. The key psychological idea implemented in ATRIUM is that people have different representations available for encoding category structure, and the goal of rapid error reduction is an important influence on the choice of representation. If rules are mentally available that can rapidly capture the structure of the categories, then they will tend to be used.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions in Automatic Control*, *19*, 716–723.
- Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*, 3–19.

- Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (1998). *The role of inattention in everyday concept learning: Identification in the service of use*. Unpublished manuscript, McMaster University, Hamilton, Ontario, Canada.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
- Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.
- Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory and Cognition, 21*, 413–423.
- Corana, A., Marchesi, M., Martini, C., & Ridella, S. (1987). Minimizing multimodal functions of continuous variables with the “simulated annealing” algorithm. *ACM Transactions on Mathematical Software, 13*, 262–280.
- Erickson, M. A. (1999). Rules and exemplars in category learning (Doctoral dissertation, Indiana University, Bloomington, 1999). *Dissertation Abstracts International, 60*(5), 2377B.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107–140.
- Goffe, W. L., Ferrier, G. D., & Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics, 60*, 65–99.
- Grings, W. W., Schell, A. M., & Carey, C. A. (1973). Verbal control of an autonomic response in a cue reversal situation. *Journal of Experimental Psychology, 99*, 215–221.
- 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., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cognitive Science, 15*, 219–250.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.
- Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1362–1377.
- 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: Academic Press.
- Kruschke, J. K. (1996). 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 *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Erlbaum.
- Lewandowsky, S., Kalish, M. L., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1666–1684.
- McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247–255.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam’s razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4*, 79–95.
- Noelle, D. C., & Cottrell, G. W. (1996). Modeling interference effects in instructed category learning. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 475–480). Hillsdale, NJ: Erlbaum.
- Noelle, D. C., & Cottrell, G. W. (2000). Individual differences in exemplar-based interference during instructed category learning. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society* (pp. 358–363). Hillsdale, NJ: Erlbaum.
- 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. (1988). 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 and Psychophysics, 45*, 279–290.
- 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 and Cognition, 22*, 352–369.
- Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of rule-described phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 375*–402.
- 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: 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., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review, 3*, 222–226.
- Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin and Review, 6*, 495–503.
- Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 119–143.