Category Generation

Alan Jern and Charles Kemp
{ajern,ckemp}@cmu.edu
Department of Psychology
Carnegie Mellon University

Abstract

People exhibit the ability to imagine new category instances and new categories, with examples ranging from everyday activities like cooking to scientific discovery. This ability, which we call *category generation*, is not addressed by standard models of category learning, which focus on classifying instances rather than generating them. We develop a probabilistic account of category generation and evaluate it using two behavioral experiments. Our results confirm that people find it natural to generate new category instances and suggest that our model accounts well for this ability.

Keywords: category learning; category generation; generative models; Bayesian modeling

Humans exhibit a wide variety of creative abilities, including the ability to imagine entirely new objects never before observed. Evolutionary biologists predict transitional species on the basis of gaps in the fossil record (e.g., *Tiktaalik*, a species with features characteristic of both aquatic and land animals); designers develop new products that combine and improve upon the strengths of existing products (e.g., the spork); professional and amateur chefs create new recipes by swapping and mixing ingredients (e.g., the Cobb salad, invented by Robert H. Cobb by combining a collection of ingredients that happened to be available in his restaurant's kitchen). Henceforth, we will refer to this capability as *category generation*. [1]

In addition to inventing new categories of objects, people create new instances of existing categories relatively commonly. While the invention of the Cobb salad might be characterized as the creation of a new category of salad, people frequently create new instances of existing salads—swapping romaine lettuce for iceburg lettuce to obtain a variation on a Caesar salad, for example. This hierarchy of category generation problems is illustrated in Figure 1. Although the figure only shows a few levels in a hierarchy, category generation could in principle take place at any level.

These examples cannot be captured by standard accounts of categorization that focus on *classification* (e.g., deciding if a new dish is a Caesar salad or a Greek salad). Whereas classification involves assigning an object to an existing category, category generation involves creating a new instance of an existing category or creating a brand new category. In this paper, we focus on one case of category generation: the generation of new instances of a category after observing examples of that category. This case is illustrated in Figure 1 by the Caesar salad branch of the hierarchy: after observing instances x_1 through x_3, a new Caesar salad, x_4, is generated. This paper explores a Bayesian approach, which proposes that categories are represented as probability distributions, and that people can generate new instances of categories by sampling from these distributions.

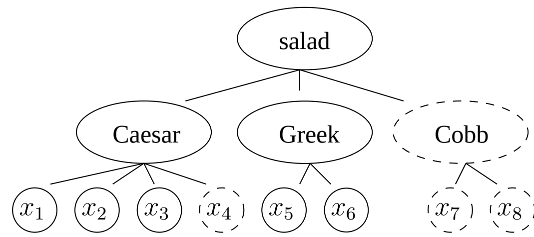

Figure 1: Category generation may take place at any level in a concept hierarchy. Two cases are illustrated here. Existing or observed knowledge is denoted by solid nodes and generated instances and categories are represented by dashed nodes. The Caesar salad branch illustrates a situation in which someone observes several instances of a Caesar salad and then generates a new instance (x_4). The Cobb salad branch illustrates the simultaneous creation of a brand new type of salad and several instances of it.

Although category generation has received relatively little attention, it has been addressed by some previous studies. Ward (1994) asked participants to invent and draw animals from a distant planet, requiring them to essentially create a new category of animal. Feldman (1997) showed people a single instance of category—a line segment with a circle on it, for example—and asked them to generate new examples of the category. Both studies confirm that people are able to generate new instances of a category, but neither provides a comprehensive formal account of this ability.

We describe a computational account of category generation that relies on Bayesian inference. Previous authors (Anderson, 1991) have developed Bayesian models of categorization, but most of these models focus on classification. Our approach uses some of the same methods as previous models, but focuses on category generation rather than classification.

We begin by reviewing some general approaches to classification, and explain why a Bayesian approach is well suited for category generation. We then describe a specific model of category generation and compare its behavior with human responses. We conclude with some general remarks about the efficacy of the Bayesian approach to category generation.

[1] The term "category generation" is sometimes used to describe tasks in which participants provide a category label, like "snacks", given instances, like "pretzels" (Ross & Murphy, 1999). The problem that we consider involves the creation of new categories or category instances, rather than the retrieval of familiar category labels.

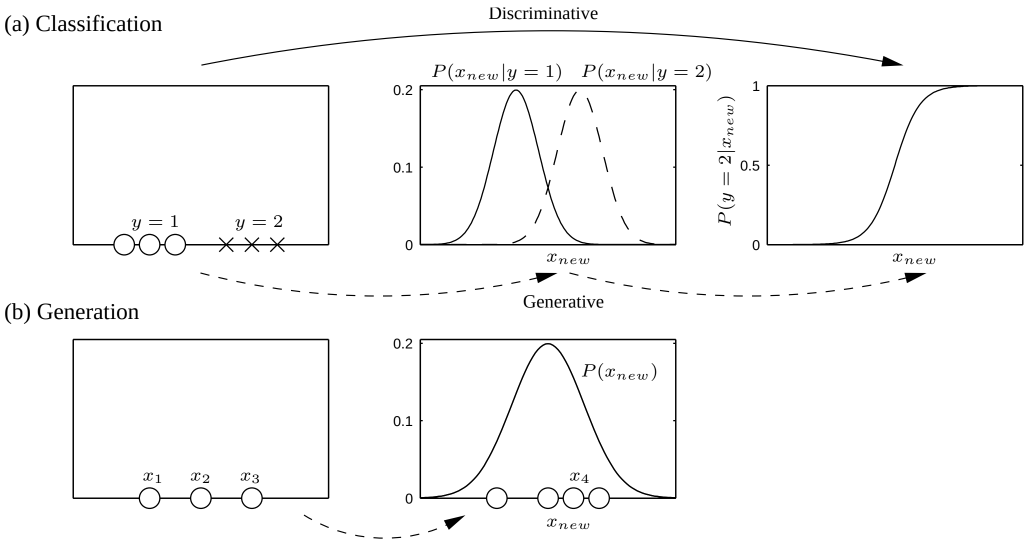

Figure 2: Discriminative classification, generative classification, and category generation. (a) Given three instances each of categories 1 and 2, a discriminative model (solid arrow) directly learns a classification distribution $P(y=2|x_{new})$ that can be used to assign category labels to new instances x_{new}. A generative model (dashed arrows) learns generation distributions $P(x_{new}|y=1)$ and $P(x_{new}|y=2)$ for each category, and these distributions induce a classification distribution via Bayes' rule. (b) Given three instances of a single category, our model learns a generation distribution $P(x_{new})$, here assumed to be Gaussian. New instances such as x_4 can then be generated by sampling from this distribution.

Classification

The standard classification problem can be formulated as follows. A set of training exemplars, $\overline{x} = \{x_1, \ldots, x_n\}$, and a corresponding set of category labels, $\overline{y}$, are provided. Each x_i is a vector of feature values. After seeing how the instances in the training set are labeled, the classification task involves assigning a category label, y_{new}, to a novel instance, x_{new}.

There are two standard approaches to classification: the *generative* approach and the *discriminative* approach. A generative model learns a probability distribution $P(x_{new}|y_{new}, \overline{x}, \overline{y})$, which we call a generation distribution, and then computes a classification distribution, $P(y_{new}|x_{new}, \overline{x}, \overline{y})$, using Bayes' rule:

$$\overbrace{P(y_{new}|x_{new}, \overline{x}, \overline{y})}^{\text{classification distribution}} \propto \overbrace{P(x_{new}|y_{new}, \overline{x}, \overline{y})}^{\text{generation distribution}} P(y_{new}|\overline{y}) \quad (1)$$

By contrast, a discriminative model learns the classification distribution directly (Bishop, 2006). The difference between the two types of models is illustrated in Figure 2a. As the figure shows, discriminative models directly learn the classification distribution, which corresponds to a soft decision boundary, while generative models begin with the intermediate step of learning the underlying distribution that generated the training data.

Most formulations of exemplar models (Nosofsky, 1985) and prototype models (Reed, 1972) are discriminative models—they can classify new instances without needing to learn the generation distribution over new instances. Anderson's (1991) rational model of categorization, however, follows a generative approach.

Our distinction between generative and discriminative approaches is standard in the machine learning literature, but terms like "generative" and "discriminative" are sometimes used differently by psychologists. Some authors reserve the term "generative" for approaches that make infinite use of finite means, and use "discriminative" to refer to settings where participants must learn to distinguish between stimuli. Note that neither usage maps perfectly onto our own.

Generative and discriminative models are both able to make predictions about human behavior on classification problems. By contrast, tasks that depend on the generation distribution, $P(x_{new}|y_{new}, \overline{x}, \overline{y})$, are naturally much better suited to a generative approach. We propose that category generation is one such task, and that learning a generation distribution allows people to generate novel instances of categories.

A Bayesian Model of Category Generation

The generation distribution, $P(x_{new}|y_{new}, \overline{x}, \overline{y})$, is defined for multiple values of y_{new} and can be used to generate instances of multiple categories. Here, however, we consider the case where there is a single category of interest. Because all exemplars have the same category label y, we drop the labels and work with the generation distribution, $P(x_{new}|\overline{x})$. Given training examples in $\overline{x}$, new examples can be generated by sampling from this distribution.

Suppose that the single category of interest is characterized by a vector of parameters $\overline{\theta}$ that is not observed. Integrating over all possible values of $\overline{\theta}$, we have

$$P(x_{new}|\overline{x}) = \int_{\overline{\theta}} P(x_{new}|\overline{\theta}) P(\overline{\theta}|\overline{x}) d\overline{\theta} \quad (2)$$

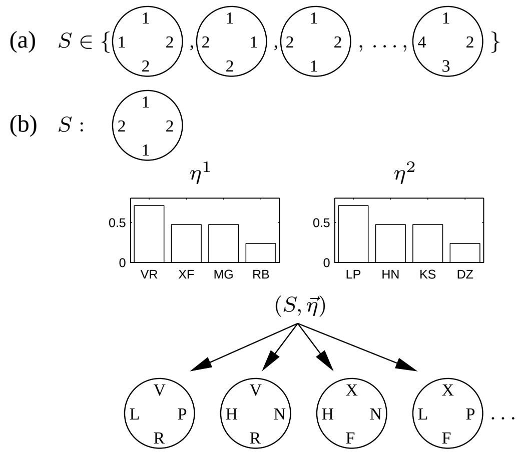

Figure 3: Stimuli for the category generation task described in the text. (a) A set of stimuli is created by first selecting a structure *S*—a partition of features into slots. The number in each feature position signifies the partition it belongs to. (b) Given *S*, the stimuli are generated by sampling from a distribution η^i over pieces for each slot *i*. Here, *S* specifies one slot made up of the top and bottom features, and one slot made up of the left and right features.

Our account of category generation is illustrated in Figure 2b for the case of a single category. Here, $\bar{\theta}$ represents the mean and variance of a Gaussian distribution. The model first infers these parameters from a set of examples and then generates new instances by sampling from that distribution. Although this procedure is simple, it cannot be carried out by a standard exemplar model, which provides a way to classify, but not generate, new instances. Note, however, that in this simple setting, new instances can be created by an approach that takes an existing exemplar and slightly varies some of its feature values. We therefore move to a richer setting where this "copy and tweak" strategy is likely inadequate.

Instead of considering cases where category instances are characterized by values along a single dimension, suppose that category instances are now represented as feature vectors. Furthermore, suppose that there are one or more latent causes that generate multiple features simultaneously, which leads to groups or clusters of features.

Here we work with a case where category instances are created by filling four locations in a circular figure with letters. Four of these instances are shown at the bottom of Figure 3b. The four locations are partitioned into one or more *slots*, and we refer to this partition as a *structure*. There are 15 possible partitions, a subset of which are shown in Figure 3a. Given the structure *S* of a category, instances of the category are created by filling each slot with a piece. Figure 3b shows a case where the structure includes a horizontal slot and a vertical slot, each of which includes two locations. The parameter $\bar{\eta}$ specifies a distribution over pieces for each slot. In Figure 3b, η^1 is a distribution over pieces that can fill the vertical slot, and η^2 is a distribution for the horizontal slot. An instance of the category can now be created by sampling a vertical piece from η^1 and a horizontal piece from η^2.

To formalize these generative assumptions, we assume that structure *S* is drawn from a uniform distribution over the 15 possible partitions, that each distribution η^i is drawn from a Dirichlet prior with parameter α, and that each piece x^i is sampled from a multinomial distribution η^i:

$$\begin{aligned} S &\sim \text{Uniform}([1:15]) \\ \eta^i|S &\sim \text{Dirichlet}(\alpha) \\ x^i|\eta^i &\sim \text{Multinomial}(\eta^i) \end{aligned} \tag{3}$$

We assume that the alphabet of symbols is fixed in advance, and that the distribution η^i is defined over all possible permutations of symbols that could fill slot *i*. For example, if the slot includes *m* cells and there are *k* symbols, then there are k^m possible pieces that could fill the slot. We set the parameter α by assuming that the prior probability that any two category instances have the same piece for a given slot is 0.5. Anderson's (1991) model of categorization makes a related assumption, and refers to the parameter 0.5 as a "coupling probability." It follows that $\alpha = \left(\frac{1}{k^m-2}, \ldots, \frac{1}{k^m-2}\right)$, where the α value for a given slot depends on the size *m* of that slot.

Now that we have formally specified our assumptions about the category we can use Equation 2 to model how novel instances of the category are generated. We set $\bar{\theta} = (S, \bar{\eta})$ and expand the second term in the integral by applying Bayes' rule:

$$\begin{aligned} P(x_{new}|\bar{x}) &= \sum_S \int_{\bar{\eta}} P(x_{new}|S,\bar{\eta}) P(S,\bar{\eta}|\bar{x}) d\bar{\eta} \\ &= \sum_S \int_{\bar{\eta}} P(x_{new}|S,\bar{\eta}) P(\bar{x}|S,\bar{\eta}) P(\bar{\eta}|S) P(S) d\bar{\eta} \end{aligned} \tag{4}$$

Each distribution on the right hand side of Equation 4 is specified by the generative assumptions in Equation 3.

Experiment 1

We designed a category generation experiment using stimuli like the circles in Figure 3 in order to test two main hypotheses: (1) that people are capable of category generation, evidenced by their ability to generate new instances of the category, and (2) that the model presented here approximates human performance on the task.

Method

Participants. Seventeen Carnegie Mellon undergraduates completed the experiment for course credit.
Design and Materials. Three different sets of of stimuli were created using the first three structures in Figure 3a, resulting in three conditions. Each participant was exposed to two of these conditions in a randomized order.

For each set, 16 different capital letters were chosen as features. All vowels were eliminated from consideration to avoid the possibility of accidental formation of pronounceable syllables or actual words. The letters, A, C, T, and G

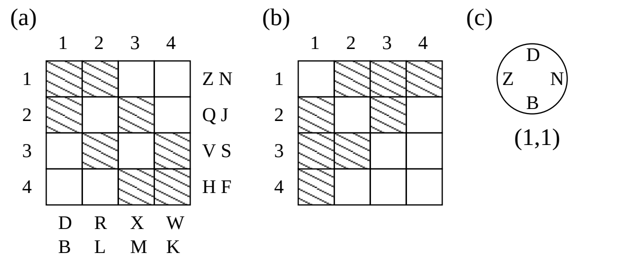

Figure 4: The stimuli used in Experiments 1 and 2. In each grid, the rows represent the possible pieces for one slot and the columns represent the possible pieces for the other slot. The rows and columns are numbered so they may be identified in the text. The hatched cells indicate which combinations were shown to participants. (a) Experiment 1 stimuli. An example set of feature values are also shown along the right and bottom edges of the grid. (b) Experiment 2 stimuli. (c) An example stimulus corresponding to item (1,1) in (a).

were also eliminated because of their semantic significance within the context of the experiment, which included a story about genomes, described below. Letters were grouped into pairs to make a total of eight pairs, four of which made up the possible values of pieces for slot 1, and the other four of which made up the possible values of pieces for slot 2. As a result, there were 16 possible combinations of pieces for each set of stimuli, of which participants were shown half. [2] The exact set of items shown to participants is indicated by the hatched cells in Figure 4a.

In addition to the training stimuli, a set of testing stimuli were prepared for a rating task. These items included some valid but unseen combinations of letter pairs (i.e. the unhatched cells in Figure 4), some seen and unseen combinations rotated 90 degrees (thus violating the structure of the category), and some distortions of seen items that matched between one and three individual features but were not consistent with the structure of the set. The rating task therefore was a typical classification task in which participants had to decide which novel items belonged in the category. The exact rating stimuli and the order in which they were presented were both randomized across participants.

Procedure. Participants were presented with the stimuli printed on index cards and were told that each item represented the genome of a strain of flu virus that had been observed in the current year. They were encouraged to spread the cards out on a table and rearrange them as they examined them. They were told that enough funds existed only to produce a flu vaccine for one additional strain of flu and were instructed to make their three best guesses of a flu virus genome that was likely to be observed but was not already in the current set. Participants made their guesses by illustrating them with a pen on paper or with a graphics tablet on the computer.

[2]Similar stimuli were used by Fiser and Aslin (2001), in which participants successfully learned to differentiate between "chunks" of symbols arranged in ambiguous grid.

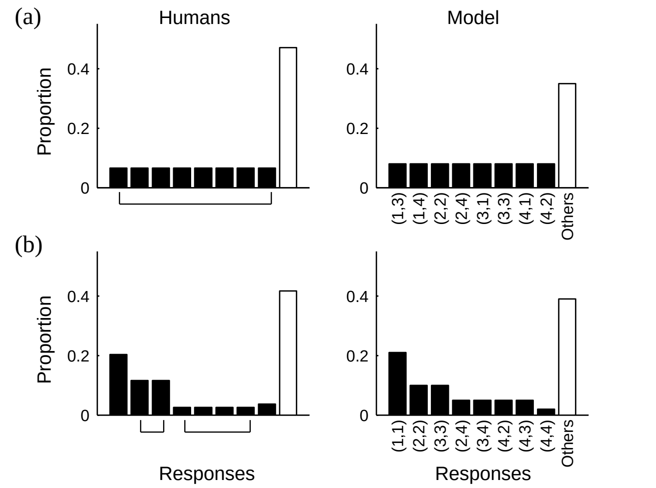

Figure 5: Comparison of human responses and model predictions for (a) Experiment 1 and (b) Experiment 2. The black bars indicate the frequency of the eight most popular responses, which are equivalent to the eight most probable responses according to the model. The white bars show the combined frequencies for all other responses. The human responses in both cases are averaged over the three conditions within the groups shown in brackets.

After making their guesses, they proceeded to a rating task in which they were shown a series of new genomes and asked to rate the likelihood (on a scale from 1 to 7) that each one represented a flu virus that would be observed this year. Thus, the first phase of the experiment was a category generation task and the second phase was a classification task.

Participants were then given a new set of cards with a different structure and repeated the preceding procedure.

Results

The model learns a category distribution that assigns nonzero probabilities to training items. To produce our predictions, we set these probabilities to 0, normalized the resulting distribution, and sampled from it.

These predictions and human responses are summarized in Figure 5a. The model predicts that the eight most probable responses correspond to the white cells in Figure 4a. These items constitute a majority (53%) of human responses. The cells in the grid are not uniquely identifiable across conditions, which used different sets of letters, so the results shown in Figure 5a are averaged across all possible alignments of cells. This averaging procedure is the reason for the remarkably uniform appearance of the behavioral data. A breakdown of responses per condition is shown in Figure 6a. Although these results are noisy, two important observations can be made. First, with the exception of structure 2, the majority of participants' responses (46% for structure 2) were valid recombinations of letter pairs. Second, among the most probable items, participants do not appear to favor any item in particular, again predicted by the model.

Due to the small training set and the highly unconstrained nature of the task, the model also predicted a fairly large number of other responses, indicated by the white bar. However, the predicted likelihoods for *individual* responses beyond the top eight are nearly negligible ($\sim 3 \times 10^{-4}$). The human responses were consistent with this prediction, and no response other than the top eight most frequent items was generated more than once.

Responses to the rating task (see Figure 7a) provide additional evidence that participants understood the structure of the category. Each participant's set of responses were converted to z-scores and then the mean scores for the different types of rating items were compared. There was a significant difference between the mean scores per participant for valid ($M = 0.64$, $SD = 0.60$) and invalid ($M = -0.26$, $SD = 0.24$) items, $t(33) = 6.26$, $p < .001$. The figure also shows mean scores for some specific types of distractors—namely, those that included between one and three previously observed pairs of features. Of particular interest are the items with three previously seen pairings (3 SP in the figure). If participants had based their judgments only on observed pairwise correlations, they would give higher ratings to the 3 SP items than the valid items, which only contain two previously seen pairings. There was a significant difference between the scores for these items ($M = -0.42$, $SD = 0.59$) and valid items, $t(33) = 6.25$, $p < .001$. These results suggest that people's responses are not primarily driven by a simple notion of feature similarity.

Taken together, our results for Experiment 1 suggest that people were able to generate new members of the category we considered, and that this ability cannot be explained by a simple similarity-based account. The two main predictions of our model were supported: people generate valid items more frequently than invalid items, but invalid items account for some proportion of responses.

Experiment 2

Although Experiment 1 provides some initial support for our model, our results are broadly consistent with an alternative model that learns rules (e.g., the rule that items are created by combining two pieces) but that does not rely on probability distributions in any fundamental way. We therefore designed a second experiment that tests the probabilistic aspect of our approach more directly. The training stimuli in Experiment 1 were created using pieces that appeared equally frequently. In Experiment 2 we replaced this balanced set of frequencies with a skewed set (see Figure 4b), and explored whether people would respond to these frequency differences as predicted by our model.

Method

The materials and procedure in Experiment 2 were identical to Experiment 1. The two experiments differed only in which set of items were shown to participants. In Experiment 2, one piece in each slot appeared three times, two pieces in each slot appeared two times, and one piece in each slot appeared once. Eighteen Carnegie Mellon undergraduates completed the experiment for course credit.

Results

The model predictions were generated the same way as in Experiment 1. The predictions and experimental results are summarized in Figure 5b. Again, not all responses were alignable across the different structures, and the averaged groups are indicated by brackets. Unlike in Experiment 1, some responses were uniquely identifiable across conditions. For example, item $(1,1)$ is the only item made of pieces that each appeared three times in the training set. Items $(2,2)$ and $(3,3)$, however, are each made up of pieces that were seen twice, and therefore must be averaged across conditions. With the exception of a small deviation from the model's prediction for the frequency of item $(4,4)$, human responses are well predicted by the model.

A breakdown of responses per condition is shown in Figure 6b. In all three cases, the most frequently generated item was the most probable item according the model. In two of the three cases, the top three most frequently generated items were the model's three most probable items. Individual responses that did not match the top eight most probable items were generated no more than twice.

Again, data from the rating task were analyzed (see Figure 7b). Two sets of ratings were excluded because the participants did not rate every item. There was a significant difference between the mean scores per participant for valid ($M = 0.55$, $SD = 0.61$) and invalid ($M = -0.22$, $SD = 0.24$) items, $t(32) = 5.23$, $p < .001$.

These results replicate our previous finding that people are able to discover the structure of a category and generate new category members that fit this structure. Our data also suggest that people are sensitive to frequency differences, a finding that is predicted by our probabilistic approach but appears less compatible with alternative rule-based accounts.

Conclusion

This paper was motivated by the observation that people are able to generate new instances of a category. Our experimental results confirmed this observation even in cases involving relatively small training sets. These results also provide support for our computational approach to category generation, which is general enough that it can be applied to many different cases of category generation.

We focused on category generation at the exemplar level, but the same basic approach may help to explain how entirely new categories are generated. For example, suppose one first learns categories that can be characterized by bivariate Gaussian categories with different means but equal covariances. Then, if asked to generate a new category in the same feature space, we might expect people to choose a new mean but preserve the covariance of the training categories. The approach presented in this paper can account for such behavior

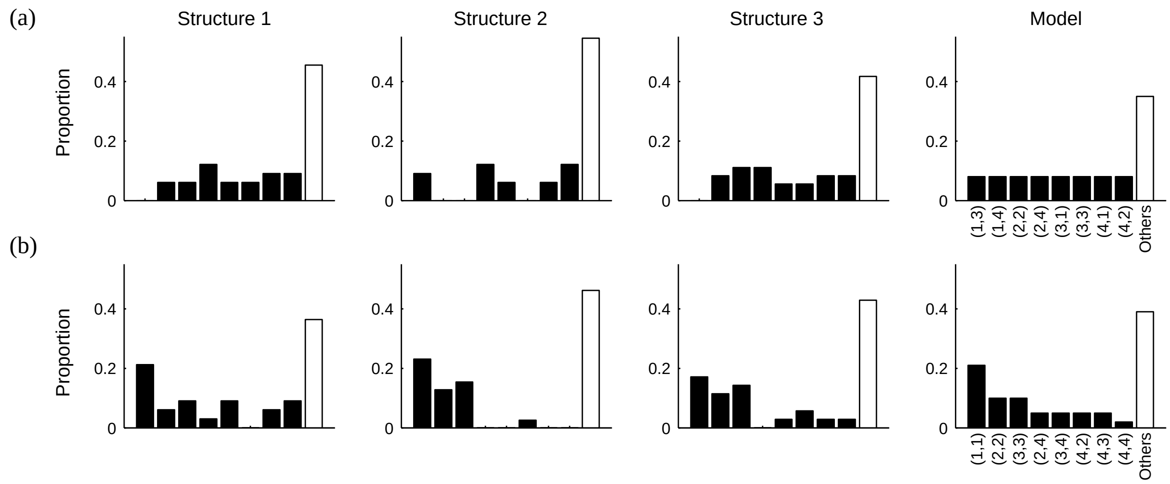

Figure 6: Comparison of human responses and model predictions for the three conditions in (a) Experiment 1 and (b) Experiment 2. In all cases, the black bars correspond to the eight most probable responses according to the model.

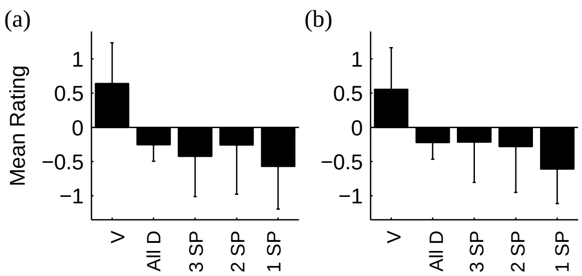

Figure 7: Mean ratings (converted to z-scores) for the test items in Experiments 1 and 2. V = Valid items; All D = All Distractors; 3 SP = Distractor with three seen pairings; 2 SP = Two seen pairings; 1 SP = One seen pairing.

with a model that learns a distribution over the means and covariances of the categories and then samples from that distribution to create a new category. It may then sample from the new category to generate instances of it (e.g., generating the first Cobb salad ever created).

Bloom (1994) has explored the hypothesis that the generative properties of natural language are inherited by other cognitive systems. Although we adopt a slightly different definition of "generative", it is clear that the ability to generate new items and ideas extends well beyond the domain of language. Consequently, the generative approach may also have applications beyond category learning—for example, to imagination and mental imagery, or to problem solving situations in which people must devise a new solution to a problem after being shown several other solutions. In the case of mental imagery, people may have some notion of a distribution over visual scenes and sample from that distribution when, say, picturing a setting described in a novel.

Although many examples of category generation (e.g., generating a new instance of a Caesar salad) seem fairly ordinary, others (e.g., inventing a Cobb salad) seem to demand more creativity. The task modeled in this paper is not especially creative, but future applications of our approach can consider tasks that require more imagination. Characterizing the computational basis of creativity is obviously a challenging problem, but a generative probabilistic approach may provide part of the solution.

Acknowledgments. We thank Faye Han for help in running the experiments and coding the data. We also thank Blair Armstrong, Michael Lee, Dan Navarro and two anonymous reviewers for helpful comments on an earlier draft of this paper.

References

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Bloom, P. (1994). Generativity within language and other cognitive domains. *Cognition*, *51*(2), 177-189.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*(2), 145-170.

Fiser, J., & Aslin, R. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499-504.

Nosofsky, R. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, *38*(5), 415-432.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 383-407.

Ross, B. H., & Murphy, G. R. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495-553.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*, 1-40.