Cognitive Psychology 66 (2013) 85–125

Contents lists available at SciVerse ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

A probabilistic account of exemplar and category generation

Alan Jern *, Charles Kemp

Carnegie Mellon University, Department of Psychology, 5000 Forbes Ave. Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Accepted 14 September 2012

Keywords:
Categorization
Generation
Sampling
Bayesian modeling

ABSTRACT

People are capable of imagining and generating new category exemplars and categories. This ability has not been addressed by previous models of categorization, most of which focus on classifying category exemplars rather than generating them. We develop a formal account of exemplar and category generation which proposes that category knowledge is represented by probability distributions over exemplars and categories, and that new exemplars and categories are generated by sampling from these distributions. This sampling account of generation is evaluated in two pairs of behavioral experiments. In the first pair of experiments, participants were asked to generate novel exemplars of a category. In the second pair of experiments, participants were asked to generate a novel category after observing exemplars from several related categories. The results suggest that generation is influenced by both structural and distributional properties of the observed categories, and we argue that our data are better explained by the sampling account than by several alternative approaches.

1. Introduction

Late one night, Robert Cobb, the owner of the Brown Derby restaurant in Hollywood, went rummaging through his restaurant's kitchen to make himself something to eat. By mixing together a handful of available ingredients including an avocado, chopped lettuce, bacon, and a hard-boiled egg, he created the Cobb salad, which can now be found in restaurants worldwide (Cobb & Willems, 1996). This event is just one demonstration of people's ability to conceive of objects and concepts that they have never encountered. Other demonstrations are found in the history of science. For example, evolutionary biologists predicted the existence of the species later named *Tiktaalik* by conceiving of

* Corresponding author. Tel.: +1 412 268 6178.
E-mail addresses: ajern@cmu.edu (A. Jern), ckemp@cmu.edu (C. Kemp).

0010-0285/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogpsych.2012.09.003

an animal that filled a gap in the fossil record between aquatic and land animals (Daeschler, Shubin, & Jenkins, 2006). Designers of all types generate objects every time they create a new sweater, light fixture, or bracelet. Even children appear to have no difficulty producing examples of new and imaginary objects (Berti & Freeman, 1997; Karmiloff-Smith, 1990).

Objects and categories can be organized into hierarchies, and in principle, new possibilities can be generated at any level of these hierarchies. Consider the simple example in Fig. 1a, in which three meals (x_1–x_3) could be described as instances of the Greek salad category or of the higher-order category of salads. We will refer to generation at the bottom level of a hierarchy as *exemplar generation*, and making a Greek salad is one example (Fig. 1c). We will refer to generation at any other level as *category generation*, and Robert Cobb's invention of the Cobb salad is one example (Fig. 1d). In this paper, we develop a probabilistic account of both exemplar and category generation. We adopt the working hypothesis that people's knowledge about categories and exemplars can be captured by probability distributions. We then propose that people generate exemplars and categories by taking advantage of one of the most basic operations supported by probability distributions: sampling. More precisely, we propose that exemplar generation amounts to sampling from a distribution over exemplars and that category generation amounts to sampling from a distribution over categories.

Our work builds on previous formal models of categorization, although most have focused primarily or exclusively on classification (Kruschke, 2008). Our computational account is most closely related to probabilistic models of classification that incorporate distributions over exemplars, such as the rational model (Anderson, 1991), the category density model (Fried & Holyoak, 1984), the mixture model of categorization (Rosseel, 2002), and the family of density estimation models described by Ashby and Alfonso-Reese (1995). Probabilistic models have also been used for related purposes such as category induction (Feldman, 1997; Tenenbaum, 1999), similarity judgement (Kemp, Bernstein, & Tenenbaum, 2005), and visual image parsing (Yuille & Kersten, 2006). Although these models have all relied on probability distributions, previous research has not highlighted the idea that new things can be generated by sampling from these distributions. Our account adopts some of the same representational assumptions as past probabilistic models, but applies these assumptions directly to the generation of exemplars and categories.

Our account can be used to develop two kinds of probabilistic models. First are models which propose that people sample from probability distributions over exemplars and categories, but make no

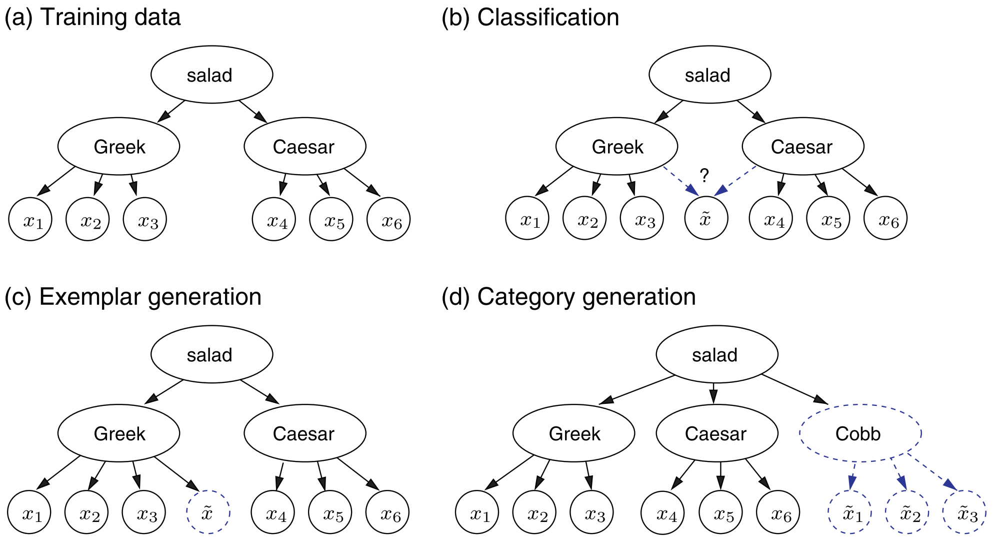

Fig. 1. Classification and generation. (a) A learner observes three Greek salads (x_1–x_3) and three Caesar salads (x_4–x_6). (b) The learner classifies a new salad $\tilde{x}$ as a Greek or Caesar salad. (c) The learner generates $\tilde{x}$, a new instance of a Greek salad. (d) The learner generates a novel category (Cobb salads) and several exemplars of this category ($\tilde{x}_1$–$\tilde{x}_3$).

strong claims about the mechanisms involved. Second are models that aim to capture the cognitive mechanisms that specify how people sample from distributions over exemplars and categories. This paper will consider both kinds of models. The mechanistic model that we evaluate proposes that a category is mentally represented as a set of distributions over the independent parts of exemplars, and that people generate novel exemplars by sampling one part at a time. Other mechanistic accounts could be considered, however, and research in this area can make contact with previous studies that invoke sampling as a cognitive mechanism. For instance, recent work has explored psychologically and neurally plausible approximations to probabilistic inference that rely on sampling (Fiser, Berkes, Orbán, & Lengyel, 2010; Hoyer & Hyvärinen, 2003; Moreno-Bote, Knill, & Pouget, 2011; Sanborn, Griffiths, & Navarro, 2010; Shelton, Bornschein, Sheikh, Berkes, & Lücke, 2011; Shi & Griffiths, 2009). Sampling has also been used to account for numerous results from the judgment and decision-making literature (Ratcliff & Smith, 2004; Stewart, Chater, & Brown, 2006; Tversky & Kahneman, 1974; Vul & Pashler, 2008).

Generation has been previously addressed by both artificial intelligence researchers and psychologists. Numerous AI models of creativity have been proposed that are broadly relevant to the problem of generation (Boden, 1998). One notable example related to our work is the Letter Spirit model of typeface design (Rehling, 2001; Rehling & Hofstadter, 2004) which generates coherent fonts of letters after observing several example letters. Another example is the model of handwritten digit recognition developed by Hinton, Osindero, and Teh (2006) which can generate digits after being trained. In the psychological literature, work on creative cognition has focused on how people create new things and what factors influence their behavior (Feldman, 1997; Marsh, Landau, & Hicks, 1996, 1999; Smith, Ward, & Schumacher, 1993). In one study, Ward (1994) asked participants to create and draw animals from another planet, and found that people's responses were influenced by properties of familiar animal categories. Previous psychological studies therefore provide strong evidence that category knowledge shapes and constrains people's behavior when generating new things. This paper argues for a similar conclusion, but goes beyond existing psychological research by providing a computational account.

2. The sampling account of exemplar and category generation

Researchers have traditionally studied people's category knowledge by asking people to classify novel exemplars. This approach has proved fruitful, but is limited by the fact that it asks people to use their category knowledge in only one particular way. An alternative approach is to examine how people use their category knowledge to address multiple problems (Kemp & Jern, 2009; Markman & Ross, 2003). For example, in addition to classification, people use their category knowledge for feature inference (Anderson & Fincham, 1996; Yamauchi & Markman, 1998), induction (Osherson, Smith, Wilkie, López, & Shafir, 1990), and causal reasoning (Waldmann & Hagmayer, 2006). A complete account of categorization should be able to account for all of the ways in which category knowledge is used. In order to evaluate the merits of competing theoretical accounts, it is therefore necessary to explore how people and models address problems other than classification (Love, Medin, & Gureckis, 2004).

One such problem that has received relatively little attention is the generation of new exemplars and categories. This paper argues that people's ability to solve generation problems places important constraints on theories of categorization. We begin in this section by presenting a *sampling account* of generation which proposes that category knowledge is represented using probability distributions and that new exemplars and categories are generated by sampling from these distributions. We also show how these probability distributions can be used to address the problem of classification. After introducing the sampling account, we consider alternative computational accounts that make different assumptions about how category knowledge is represented and used.

2.1. Classification

Imagine that you have seen the six examples of Greek and Caesar salads in Fig. 1a. A friend asks you to get a Greek salad from a display case containing different salads. Deciding what salads in the case do and do not belong to the category of Greek salads is an example of a typical classification problem.

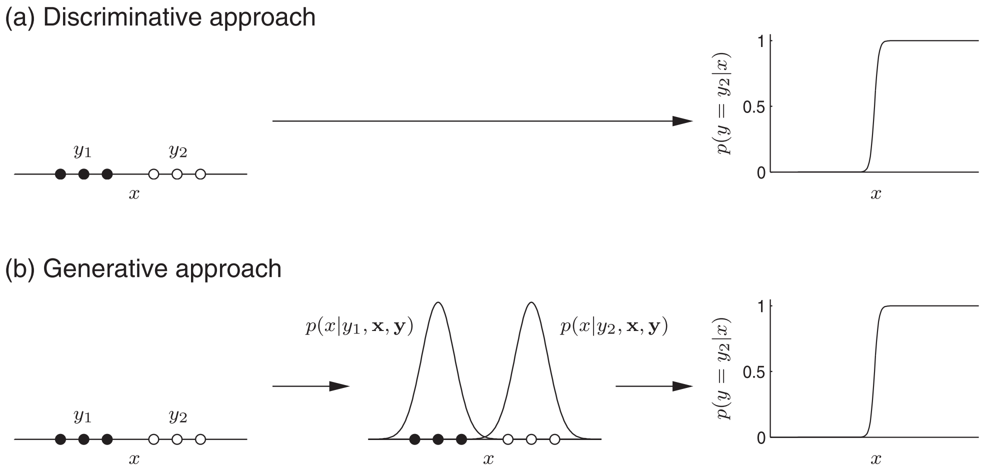

Fig. 2. The discriminative and generative approaches to classification. In both cases, three exemplars of categories y_1 and y_2 are given. (a) A discriminative model directly learns the classification distribution, $p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y})$, which has been plotted for y_2. This distribution can be used to assign category labels to novel exemplars $\tilde{x}$. (b) A generative model first learns a joint distribution over feature vectors and category labels, $p(\tilde{x}, \tilde{y}|\mathbf{x}, \mathbf{y})$, that is likely to have generated the training data. (For clarity, the so-called class-conditional distributions $p(\tilde{x}|\tilde{y}, \mathbf{x}, \mathbf{y})$ are shown instead of the full joint distribution.) The joint distribution can then be used to compute the classification distribution as shown in the text.

To formalize this problem, let $\mathbf{x} = \{x_1, \ldots, x_6\}$ be the set of training exemplars (the observed salads), where each x_i is a vector of feature values (in this case, a list of ingredients). Let $\mathbf{y}$ be a vector of category labels such that y_i is the category label for x_i, where $y_i \in$ {Greek salad, Caesar salad}. After exposure to the training exemplars, the subsequent classification problem is depicted in Fig. 1b. In the figure, $\tilde{x}$ represents a novel salad, and you must infer whether this salad belongs with the Greek salads or the Caesar salads. This inference corresponds graphically to inferring which of the two dashed arrows in the hierarchy is appropriate. In all of the panels of Fig. 1 we use solid lines to represent information that is given to a learner and dashed lines to represent information that must be supplied by the learner.

The problem of classification can be formulated using the probability distribution $p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y})$, which specifies the extent to which category label $\tilde{y}$ is appropriate for exemplar $\tilde{x}$.[1] We will refer to this distribution as the *classification distribution*. There are two distinct approaches to learning the classification distribution: the *generative* approach and the *discriminative* approach (Bishop, 2006; Ng & Jordan, 2002). A model that uses the generative approach first learns a joint distribution over feature vectors and category labels, $p(\tilde{x}, \tilde{y}|\mathbf{x}, \mathbf{y})$. The classification distribution can then be computed as follows:

$$p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y}) = \frac{p(\tilde{x}, \tilde{y}|\mathbf{x}, \mathbf{y})}{p(\tilde{x}|\mathbf{x}, \mathbf{y})} = \frac{p(\tilde{x}, \tilde{y}|\mathbf{x}, \mathbf{y})}{\sum_{\tilde{y}} p(\tilde{x}, \tilde{y}|\mathbf{x}, \mathbf{y})}. \tag{1}$$

By contrast, a model that uses the discriminative approach learns the classification distribution directly. This difference is illustrated in Fig. 2, where both approaches are applied to the same classification problem. Fig. 2a shows one example of a discriminative model that assumes a probabilistic decision boundary between two categories. This boundary takes the form of an “S-shaped” curve called a logistic function. The model adjusts the parameters of this function, like its position along the x-axis and how steeply it rises, to best fit the training data. Fig. 2b shows one example of a generative model that assumes the two categories are normally distributed. Based on the training data,

[1] We characterize the classification problem using a probability distribution to be as general as possible. Note, however, that a solution to a classification problem may be entirely deterministic (i.e., a function that maps feature vectors $\tilde{x}$ to category labels $\tilde{y}$). This is simply a special case of our formulation in which $p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y}) \in \{0, 1\}$ for all $\tilde{x}$.

the model learns the base rates $p(\tilde{y}|\mathbf{x},\mathbf{y})$ of the two categories, and the means and variances that specify their respective distributions, $p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y})$. Learning these two sets of distributions is the same as learning the joint distribution over feature vectors and category labels, which can be rewritten as the product $p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y})p(\tilde{y}|\mathbf{x},\mathbf{y})$.

2.2. Exemplar generation

Observing the six salads in the training set should allow you to solve other problems in addition to classification. For example, suppose that your friend asks you to *make* a Greek salad. This problem, which we refer to as exemplar generation, is depicted in Fig. 1c. Here, $\tilde{x}$ represents an exemplar that must be generated rather than classified. Consistent with our graphical notation, $\tilde{x}$ is shown as a dashed circle, but the arrow that connects $\tilde{x}$ to the Greek salad category is solid to indicate that you were asked to generate an exemplar of that category.

The problem of exemplar generation can be formulated in terms of learning the probability distribution $p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y})$, which we will refer to as the *exemplar generation distribution*. This distribution can be used in cases where a category label $\tilde{y}$ is given and an exemplar $\tilde{x}$ from that category must be generated. After learning the joint distribution $p(\tilde{x},\tilde{y}|\mathbf{x},\mathbf{y})$ over feature vectors and category labels, the generative approach can compute the exemplar generation distribution as follows:

$$p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y}) = \frac{p(\tilde{x},\tilde{y}|\mathbf{x},\mathbf{y})}{p(\tilde{y}|\mathbf{x},\mathbf{y})} = \frac{p(\tilde{x},\tilde{y}|\mathbf{x},\mathbf{y})}{\sum_{\tilde{x}} p(\tilde{x},\tilde{y}|\mathbf{x},\mathbf{y})}. \qquad (2)$$

Note that this computation is the same as Eq. (1) for the classification distribution, except that the roles of $\tilde{x}$ and $\tilde{y}$ are reversed.[2] The correspondence between the two equations indicates that the generative approach is equally well-suited for classification and exemplar generation. In contrast, the discriminative approach learns only the classification distribution, and therefore has no principled way of deriving the probability distribution required for exemplar generation.

Exemplar generation can be viewed as the opposite of classification. In classification, $\tilde{x}$ is given and a learner must provide its category label. In exemplar generation, a category label is given and a learner must generate a corresponding exemplar $\tilde{x}$. A critical difference between these two problems is the amount of information that the learner must produce: classification requires the learner to produce only a single category label, but generation requires the learner to produce a complete set of exemplar features. One consequence of this difference is that exemplar generation demands a richer category representation than might be necessary for classification. For example, suppose that all Greek salads contain olives and all Caesar salads do not. Classifying a new salad then requires attending only to the presence or absence of olives. Generating a new Greek salad, however, requires attending to more than just the diagnostic dimensions, as a bowl of olives would be a poor instance of a Greek salad.

Exemplar generation can also be viewed as an extreme form of feature inference. A typical feature inference problem requires a learner to predict unobserved features of a category exemplar. For example, if a friend points at a distant salad and calls it a Greek salad, you might infer that it contains olives even if you are unable to see any olives. Generation can be formulated as a problem in which none of the features of the exemplar in question are observed and all of them must be simultaneously predicted. Previous studies have established that feature inference often draws on richer category knowledge than classification (Yamauchi & Markman, 1998, 2000), and generation is likely to be even more demanding than feature inference. Consider, for example, the difference between classifying a poem as a sonnet, filling in the missing final word of a sonnet, and generating an entirely new sonnet.

We propose that solving an exemplar generation problem requires two steps. The first step is to use a set of training exemplars to learn the properties of a category. In the case of the Greek salad, a learner might observe that the ingredients common to all the training exemplars are tomatoes, cucumbers, and olives, and that none of the training exemplars contain lettuce. The second step is to generate a new exemplar using the learned category properties. Both steps are naturally captured using

[2] Analogously, we assume that $\tilde{x}$ is discrete. If this were not the case, the summation in the denominator of the rightmost expression would be replaced with an integral.

probability distributions. For salads, the first step is to use the training set to learn a probability distribution over a space of ingredients for Greek salads. This distribution would likely be different than the one for Caesar salads. The second step is to sample an exemplar, a Greek salad, from the learned distribution. These two steps are shown for a one-dimensional, normally distributed category in Fig. 3a.

The sampling account formalizes these steps as follows. We first assume that each category y is characterized by a vector of unobserved parameters $\boldsymbol{\theta}_y$. Suppose that we wish to generate an exemplar from category $\tilde{y}$. Because we assume that all relevant information about $\tilde{y}$ is contained in the vector of parameters $\boldsymbol{\theta}_{\tilde{y}}$, the exemplar generation distribution can be rewritten as $p(\tilde{x}|\boldsymbol{\theta}_{\tilde{y}})$. New exemplars may then be generated by sampling from this distribution. This procedure is illustrated in Fig. 3a for a normally distributed category. Here the category parameters are $\boldsymbol{\theta}_{\tilde{y}} = (\tilde{\mu}, \tilde{\sigma}^2)$, the mean and variance of the category. The model first infers these parameters from the training data, leading to an estimate of the exemplar generation distribution: a normal distribution with the inferred mean and variance. Next, the model generates a new exemplar $\tilde{x}$ by sampling from this distribution. This procedure can be repeated to generate multiple exemplars. Section A.1 of the Appendix describes how the same basic idea can be implemented by averaging over all possible values of the category parameters instead of choosing the single best estimate of these parameters.

2.3. Category generation

After observing examples of many different kinds of salads (Caesar salads, Greek salads, etc.), you would likely notice general properties that characterize the higher-order category of salads. This knowledge might allow you to generate an entirely new salad category, just as Robert Cobb did when he created the Cobb salad category. The problem of category generation is represented using our graphical notation in Fig. 1d. Here, the goal is to generate a new salad category. Therefore the arrow connecting the higher-order salad category to the Cobb salad category is solid. The dashed circles and arrows that make up the Cobb salad category and its exemplars indicate that these elements were generated.

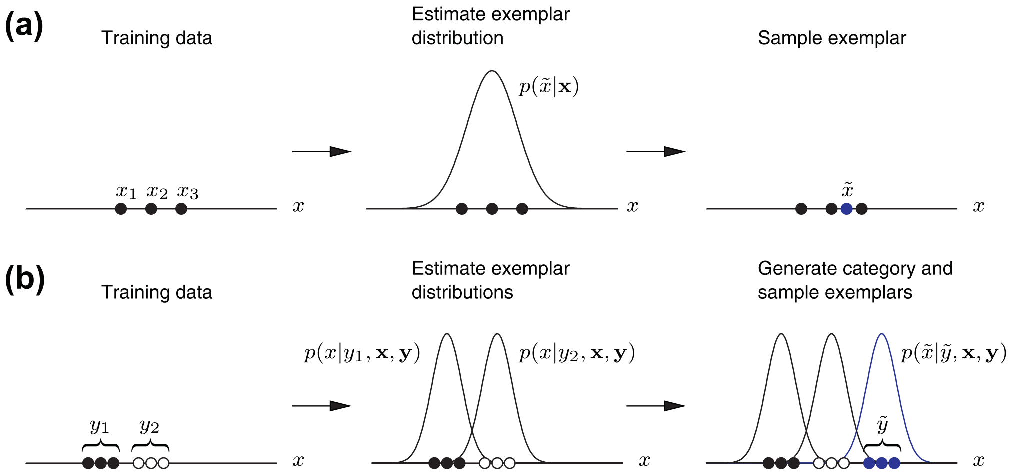
(a)
Training data
x_1 x_2 x_3
x
Estimate exemplar distribution
$p(\tilde{x}|\mathbf{x})$
x
Sample exemplar
$\tilde{x}$
x
(b)
Training data
y_1 y_2
x
Estimate exemplar distributions
$p(x|y_1, \mathbf{x}, \mathbf{y})$
$p(x|y_2, \mathbf{x}, \mathbf{y})$
x
Generate category and sample exemplars
$p(\tilde{x}|\tilde{y}, \mathbf{x}, \mathbf{y})$
$\tilde{y}$
x

Fig. 3. The sampling account of exemplar and category generation. (a) Exemplar generation first involves estimating a distribution over exemplars from the training data. For a normally distributed category, this step requires estimating the mean and variance of the category. A new exemplar $\tilde{x}$ is then generated by sampling from the estimated distribution. (b) Category generation first involves estimating distributions over exemplars for each training category (here y_1 and y_2) and determining what these categories have in common. Here the categories are both normally distributed with different means but equal variances. A new category $\tilde{y}$ is generated that has the same variance as the training categories.

The problem of category generation can be formulated using the probability distribution

$$p(\tilde{x}, \boldsymbol{\theta}_{\tilde{y}} | \mathbf{x}, \mathbf{y}) = p(\tilde{x} | \boldsymbol{\theta}_{\tilde{y}}) p(\boldsymbol{\theta}_{\tilde{y}} | \mathbf{x}, \mathbf{y}), \tag{3}$$

where $\boldsymbol{\theta}_{\tilde{y}}$ denotes the set of parameters that characterize the generated category $\tilde{y}$. As the equation shows, the probability distribution required for category generation can be decomposed into a product of two distributions. On the right is the distribution $p(\boldsymbol{\theta}_{\tilde{y}} | \mathbf{x}, \mathbf{y})$, which we will refer to as the *category generation distribution*. On the left is the distribution $p(\tilde{x} | \boldsymbol{\theta}_{\tilde{y}})$, the exemplar generation distribution for the generated category $\tilde{y}$. This decomposition highlights the fact that category generation involves first generating a set of category parameters $\boldsymbol{\theta}_{\tilde{y}}$ and then generating a set of exemplars from the resulting category.

Generating a category like a Cobb salad requires knowledge about the properties that related categories share. Just as a bowl of olives would be a poor instance of a Greek salad, the set of your cousin's favorite colors would be a poor proposal for a new salad category. In other words, a new category must be not only coherent, but should also respect the hierarchy within which it is situated. This perspective suggests that category generation consists of four steps. The first two steps are to learn the properties that characterize each training category (e.g., Greek salads and Caesar salads) and the characteristics that these categories share (e.g., the unifying characteristics of salad categories). The third step is to generate a new category (e.g., the Cobb salad category) that shares the unifying characteristics identified in the second step. The fourth and final step is to generate one or more exemplars of the new category (e.g., one or more Cobb salads). Fig. 3b illustrates this procedure for a case with two normally distributed training categories with uniformly distributed means but approximately equal variances.

In order to learn the characteristics that a set of training categories share, we treat these categories as exemplars of a higher-order category, which is itself characterized by a vector of parameters. More precisely, we assume that the parameters $\boldsymbol{\theta}_y$ that characterize each category y were sampled from a higher-order category distribution with its own vector of parameters, which we will denote by $\boldsymbol{\phi}$. For the example in Fig. 3b, each category y can be characterized by its mean and variance, that is, $\boldsymbol{\theta}_y = \left(\mu_y, \sigma_y^2\right)$. The higher-order category can be characterized by the mean variance, that is, $\boldsymbol{\phi} = \sigma_0^2$. A generative model would first infer the values of the category parameters $\boldsymbol{\theta}_y$ and the higher-order category parameters $\boldsymbol{\phi}$ from the training data, leading to an estimate of the category generation distribution: a probability distribution over variances with mean σ_0^2. A new category $\tilde{y}$ can then be generated by sampling a variance from this distribution and sampling a mean from a uniform distribution. Exemplars of this new category can be generated by sampling from a normal distribution with these new parameters. This procedure can be repeated to generate multiple novel categories.

Learning what unifies a family of categories has previously been explored with children (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002) and adults (Perfors & Tenenbaum, 2009). One developmental study suggests that children learn to expect exemplars of an object category to be similar in shape, even if they vary along other dimensions such as color and texture (Smith et al., 2002). The tendency to generalize on the basis of shape is sometimes referred to as a shape bias (Landau, Smith, & Jones, 1988), and subsequent work has shown similar biases in children's generalizations (Colunga & Smith, 2008; Jones, Smith, & Landau, 1991; Keil, Smith, Simons, & Levin, 1998; Macario, 1991). These studies suggest that people are able to recognize properties shared by entire classes of categories, and we propose in addition that people are able to use this knowledge to generate novel categories.

2.4. A mechanistic sample-by-parts account

The sampling account proposes that people are able to sample from probability distributions, but makes no specific claims about the mechanisms involved. In some contexts, sampling can be plausibly assumed to be a primitive cognitive mechanism. For example, it is possible that a single primitive mechanism allows memory traces to be sampled (Tversky & Kahneman, 1974; Stewart et al., 2006; Vul & Pashler, 2008). The sampling account, however, relies on distributions over exemplars and categories that may be relatively complex, and sampling from these distributions may rely on more basic cognitive mechanisms.

There may be many ways to produce samples from the same probability distribution. For example, one way to generate from an exemplar distribution is to sample an entire exemplar in a single operation. An alternative is to sample the parts of an exemplar in turn. For example, suppose that a salad corresponds to a combination of three parts: a green, a vegetable, and a protein. Sampling from a distribution over salads can then be achieved by a process akin to rolling a set of weighted dice. An arugula-squash-tofu salad could be generated by rolling three dice in order to sample from distributions over greens, vegetables, and proteins, respectively. We refer to this general approach as the *sample-by-parts* account, and will evaluate it using data from our first two experiments. Cognitive mechanisms that correspond to rolling mental dice have been previously considered. For example, any process-level model that relies on the Luce choice rule (Luce, 1959) must invoke a mechanism of this kind.

Sampling from exemplar and category distributions may be supported by different primitive mechanisms in different contexts, but sample-by-parts is one of the simplest possible accounts of exemplar generation. For categories with correlated parts, other sampling mechanisms will be needed. Characterizing the full set of sampling mechanisms is a challenge for future work, but our discussion of sample-by-parts will serve to illustrate how the sampling account can be given a mechanistic grounding.

3. Alternative accounts of generation

The previous section described how the generative approach to categorization can account for both classification and generation. We now discuss alternative theoretical accounts of categorization and consider the extent to which they can account for generation. Many different models of categorization have been proposed, but most of them are fundamentally discriminative in nature. For example, the GCM (Nosofsky, 1986), ALCOVE (Kruschke, 1992), and SUSTAIN (Love et al., 2004) are three well-known examples of discriminative models and there are many others (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Pothos & Chater, 2002; Rehder & Murphy, 2003; for reviews, see Kruschke, 2008; Pothos & Wills, 2011; Wills & Pothos, 2012). We suggested above that discriminative models have no intrinsic way of generating new exemplars and categories. Now we consider how these models might be extended in order to account for generation.

As described earlier, a discriminative model provides a classification distribution that can be used to classify novel exemplars. In order to transform a discriminative model into a model of generation, this classification distribution must be supplemented with an additional cognitive mechanism. Choosing a simple mechanism ensures that the resulting approach will not depart too far from the discriminative approach to categorization. A later section evaluates an approach that supplements the classification function with a mechanism that samples randomly from the full space of exemplars. This mechanism can be used to develop a *randomly-sample-and-score* account that samples a set of exemplars, scores each one using the classification distribution, and then retains only the highest-scoring exemplars. Our analyses suggest, however, that randomly-sample-and-score models are not psychologically plausible because they require a vast number of candidate exemplars to be sampled in order to identify the handful of best-scoring exemplars that end up being generated.

An alternative approach is to supplement the classification distribution with a heuristic-based search algorithm that can efficiently identify the best-scoring exemplars. This search algorithm, however, will necessarily be more complex than the additional mechanism assumed by the randomly-sample-and-score account, perhaps considerably so. As a result, the specification of the search algorithm will need to draw on principles that go beyond the principles captured by discriminative accounts of classification. In contrast, the sampling account provides a parsimonious account of categorization that uses a single set of principles to account for both classification and generation.

The approaches just described can be used to convert any discriminative model into an account of generation. Other possible accounts of generation might be developed by taking advantage of the unique features of specific discriminative models. We now discuss several popular discriminative approaches and consider how each one might be extended to account for generation. We focus on three model classes: exemplar models, decision bound models, and rule-based models.

3.1. Exemplar models

Exemplar models like the GCM (Nosofsky, 1986) and ALCOVE (Kruschke, 1992) classify novel exemplars by computing their similarity to previously observed exemplars stored in memory. These models can be extended to account for generation by assuming that stored exemplars can be copied and changed in some way in order to produce novel exemplars. We will refer to the resulting approach as the *copy-and-tweak* account, and the same basic approach has been proposed by several previous authors (Barsalou & Prinz, 1997; Barsalou, 1999; Ward, 1995; Ward, Patterson, Sifonis, Dodds, & Saunders, 2002; Ward, Patterson, & Sifonis, 2004).

Although the copy-and-tweak account may seem conceptually different from the sampling account, in some cases the two approaches are equivalent. Consider, for example, a problem in which the exemplars are points in n-dimensional space. A copy-and-tweak account would retrieve one of these exemplars and change it slightly along each dimension. This procedure is conceptually identical to sampling from a sum of normal distributions centered on each observed exemplar, which is formally equivalent to the GCM (Ashby & Alfonso-Reese, 1995; Jäkel, Schölkopf, & Wichmann, 2008). Thus, in many cases, applying copy-and-tweak to an exemplar model will be equivalent to sampling from the generation distribution learned by a generative model. The relationship between copy-and-tweak and the sampling account suggests that both accounts will make similar predictions for problems in which people generate from categories that correspond to clusters in a similarity space. The two accounts, however, will diverge in cases where category membership cannot naturally be captured by similarity. Our first set of experiments uses structured categories that are specifically designed to distinguish between the two accounts.

3.2. Decision bound models

Decision bound models learn a boundary in the feature space that separates exemplars of different categories (Ashby & Gott, 1988; Ashby & Maddox, 1992). For example, the classification distribution in Fig. 2a is a soft decision boundary that separates exemplars of two categories. Decision bound models are perhaps the purest examples of discriminative models, because they do not maintain representations of individual exemplars or categories. The fact that exemplars are not stored means that the copy-and-tweak approach just described cannot be applied.

In order to account for generation, a decision bound model could be supplemented with a mechanism that uses the decision boundary in some way. For example, in order to generate from category y, a model could sample a random point along the decision boundary and then move away from this point in the direction corresponding to category y. Without additional assumptions, however, the resulting approach would not be sensitive to distributional information. For example, it would not capture the idea that some exemplars of category y occur more frequently than others, and we show later that people use this information when generating new exemplars.

3.3. Rule-based models

Rule-based models like RULEX (Nosofsky, Palmeri, & McKinley, 1994; Nosofsky & Palmeri, 1998) learn verbally specifiable rules that define different categories. In some cases, these rules could serve as templates for generating new exemplars. For instance, RULEX learns category representations that maintain "wildcard" features that are unnecessary or uninformative for distinguishing between contrasting categories. RULEX could therefore be used to generate new exemplars of a category by starting with a learned category template and filling the wildcard features with randomly sampled values. Once again, however, this approach would need additional assumptions in order to account for distributional information.

3.4. Models evaluated in this paper

Because additional theoretical work would be needed in order to turn decision bound models and rule-based models into viable accounts of generation, we will not consider these models any further. We focus instead on three accounts of generation: the sampling account, the randomly-sample-and-score account, and the copy-and-tweak account. We believe that these three accounts represent the simplest possible ways to turn existing accounts of classification into accounts of generation. As already described, the sampling account builds on generative models of classification, the randomly-sample-and-score account builds on discriminative models of classification, and the copy-and-tweak account builds specifically on exemplar models of classification. Future researchers may wish to develop and evaluate alternative accounts of generation, but the three considered here will serve to illustrate how the problem of generation can inform theoretical accounts of categorization.

4. Experiment 1: exemplar generation

We now evaluate the three accounts just described by comparing them to human behavior in two experiments that focus on exemplar generation. Later we describe two experiments that focus on category generation. Our initial presentation of the exemplar generation experiments is organized around the sampling account of generation. After presenting the results of Experiment 2, we return to the randomly-sample-and-score and copy-and-tweak approaches, and discuss the extent to which they can account for Experiments 1 and 2.

In each exemplar generation experiment, we present participants with a set of observed category exemplars and ask them to generate novel exemplars. Given that this task is relatively unconstrained, our first and most basic question is whether participants tend to generate the same exemplars. If consistent patterns emerge, then our next question is how well the three theoretical accounts can explain these patterns. Before introducing the specific predictions for Experiment 1, we describe the category used in both experiments.

4.1. An exemplar generation task

We designed a category in which sets of features appear in correlated groups according to a latent category structure or schema. Examples of our stimuli are shown at the bottom of Fig. 4b. Category exemplars are generated by filling four feature positions in a circle with capital letters. The feature positions are partitioned into one or more *slots*, and we refer to a partition of positions into slots as a *structure*. There are 15 possible structures, four of which are shown in Fig. 4a. After choosing a structure S for a category, exemplars of that category are generated by filling each slot with a *part*, which is composed of letters. Fig. 4b shows a case where the structure consists of a horizontal slot and a vertical slot, each of which includes two positions. Next, a distribution over possible parts for each slot is specified. Fig. 4b shows distributions over parts that can fill Slot 1 (the vertical slot) and Slot 2 (the horizontal slot). An exemplar of the category can now be generated by sampling a vertical part from the Slot 1 distribution and a horizontal part from the Slot 2 distribution.

In our exemplar generation task, participants were given a set of stimuli generated from a structure with two slots and then asked to generate novel exemplars of the category. We chose training sets that included only a subset of possible combinations of parts, allowing participants to generate novel exemplars by combining parts. The stimuli used in our task are similar in spirit to those used by Fiser and Aslin (2001), who found that participants successfully learned to organize grids of symbols into recurring “chunks.” Their study, however, focused on classifying novel exemplars whereas our task focuses on generating exemplars.

The sampling account can accommodate structured categories that are characterized by distributions over parts. When applied to our experiments, the sampling account typically generates novel exemplars by combining parts that have not previously been paired. For example, after observing the four exemplars in Fig. 4b, the sampling account might generate a novel exemplar that combines a vertical part (D B) with a horizontal part (e.g., V S). As described in a later section, the sampling and copy-and-tweak accounts make different predictions. Experiment 1 therefore aims to distinguish between these accounts by evaluating the prediction that participants will consistently generate novel exemplars by combining previously observed parts.

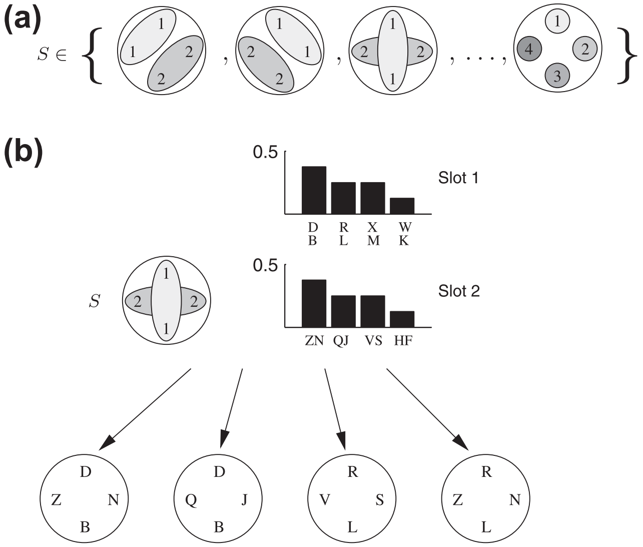

Fig. 4. Stimuli for the exemplar generation task used in Experiments 1 and 2. (a) A set of stimuli is generated by first selecting a structure *S*, a partition of features into slots. The number in each feature position denotes the partition that it belongs to. The partitions are also illustrated by the shaded ovals, but these ovals were not present in the stimuli shown to participants. (b) Given *S*, exemplars are generated by sampling from a distribution over parts for each slot. Here, *S* specifies one slot made up of the top and bottom positions, and one slot made up of the left and right positions. The figures at the bottom are examples of the stimuli that participants actually saw.

Experiment 1 also evaluates a second prediction that is specifically motivated by the mechanistic sample-by-parts account. Sample-by-parts predicts not only that participants will generate exemplars that correspond to combinations of observed parts, but that participants will generate these exemplars one part at a time. For example, after observing the exemplars in Fig. 4b, a sample-by-parts model might generate a novel exemplar by sampling a vertical part and then sampling a horizontal part. Other strategies are possible: for example, a participant might generate the four letters in clockwise order. To evaluate our second prediction, we recorded the order in which participants wrote the letters in each novel exemplar, and we will assess whether people tended to write letters belonging to the same part in sequence.

4.2. The structured sampling model

We now describe how the sampling account can be applied to exemplar generation tasks involving the structured category in Fig. 4b. The task can be formulated using our earlier notation. The choice of structure and slot distributions define the training category $\tilde{y}$, the exemplars of the category given to the learner comprise the training data $\mathbf{x}$ (and their corresponding category labels $\mathbf{y} = \{\tilde{y}, \ldots, \tilde{y}\}$), and the learner generates new category exemplars $\tilde{\mathbf{x}}$. According to the sampling account, these exemplars are drawn from the exemplar generation distribution $p(\tilde{x}|\tilde{y}, \mathbf{x}, \mathbf{y})$. In order to estimate this distribution, we must first infer the category parameters $\boldsymbol{\theta}_{\tilde{y}}$, which in our task include the structure and slot distributions. Let η_i specify the distribution over parts for slot i. Then, $\boldsymbol{\theta}_{\tilde{y}} = (S, \eta_1, \ldots, \eta_{|S|})$, where $|S|$ is the number of slots in structure S. Because this model incorporates a representation of the structure of the category, we refer to it as the structured sampling model. Appendix A shows how we estimate $\boldsymbol{\theta}_{\tilde{y}}$, and includes complete specifications of all other models described in this article. After the model infers $\boldsymbol{\theta}_{\tilde{y}} = (S, \eta_1, \ldots, \eta_{|S|})$, it can use these parameters to generate new exemplars by sampling from the exemplar generation distribution, $p(\tilde{x}|S, \eta_1, \ldots, \eta_{|S|})$.

4.2.1. Sampling-based comparison models

Although the structured sampling model does not assume much knowledge about the relative probabilities of structures or the distributions over parts, it does assume that the structure S and the slot distributions both exist. It is natural to ask whether people actually learn both of these elements and take them into account when generating exemplars. To address this question, we evaluate the predictions of two additional sampling-based models that are created by subtracting elements from the structured sampling model. We refer to these models as the rule-based model and the independent features model.

Rule-based model. The rule-based model can learn the structure of the category but is unable to learn graded probability distributions over the slots. Instead, the model simply assumes that a part can appear in a slot only if it has been previously observed in that slot. This model is equivalent to the structured sampling model if the slot distributions are all assumed to be uniform distributions over the parts that have been observed in each slot. We refer to this model as the rule-based model because the information that it learns can be viewed as a set of logical rules that specify which slots exist and which pieces can fill these slots.

Independent features model. The independent features model can learn graded probability distributions over the features that fill each feature position but does not learn the underlying structure S. This model is equivalent to the structured sampling model if the structure S is assumed to consist of four independent features. The independent features model can learn which features tend to appear in which positions, but does not consider how multiple features are related to one another.

4.3. Method

4.3.1. Participants

Thirty Carnegie Mellon University undergraduates completed the experiment for course credit.

4.3.2. Design and materials

We created three different sets of stimuli based on the first three structures in Fig. 4a, 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. Vowels (including Y) were excluded to prevent the formation of words and pronounceable syllables. The letters C, T, and G were also excluded because of their semantic significance within the context of the experiment, which included a story about genomes. This left 17 letters, of which one letter was randomly excluded for each set. Letters were randomly grouped into pairs to make a total of eight pairs, four of which made up parts for Slot 1, and the other four of which made up the parts for Slot 2. As a result, there were 16 possible combinations of parts for each set of stimuli, of which participants were shown half. The exemplars participants observed are indicated by the shaded cells of the grid in Fig. 5a, although the letters varied across conditions. As an example, the shaded cell (1,1) indicates that participants saw an exemplar in which Slot 1 was filled with the letter pair ZN and Slot 2 was filled with the letter pair DB. The resulting stimulus is shown in Fig. 5c. Note that participants saw each part exactly two

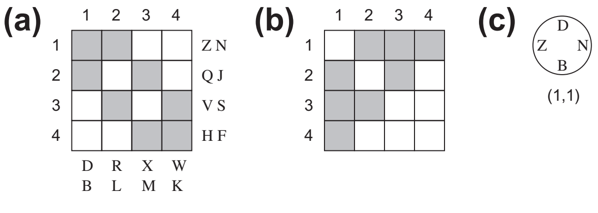

Fig. 5. The stimuli used in Experiments 1 and 2. In each grid, the rows represent the possible parts for one slot and the columns represent the possible parts for the other slot. The rows and columns are numbered for ease of identification in the text. The shaded cells indicate which combinations were shown to participants. (a) Experiment 1 stimuli. A representative set of parts is also shown along the right and bottom edges of the grid. (b) Experiment 2 stimuli. (c) An example stimulus corresponding to exemplar (1,1) from panel (a).

times. In Experiment 2, we used a different training set in which the parts did not appear with equal frequency.

In addition to the training stimuli, we also created a set of testing stimuli for a rating task. This set included some unobserved combinations of parts (i.e., the unshaded cells in Fig. 5), some observed and unobserved combinations of parts rotated 90° (thus violating the structure of the category), and some distortions of observed exemplars that contained between one and three pairs of features that appeared in the training set but were not consistent with the structure of the set. The rating task was a typical classification task in which participants had to decide which novel exemplars belonged in the category.

4.3.3. 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. The cards remained on the table throughout the experiment in order to reduce memory demands. Participants were told that there was only enough funding to produce a vaccine for one additional strain of flu and were asked 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. Some participants illustrated their guesses with a graphics tablet on the computer. We had a limited number of tablets, so the remaining participants illustrated their guesses with a pen on paper. Participants were randomly assigned to tablet or paper.

After making their guesses, participants 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. The exact rating stimuli and the order in which they were presented were both randomized across participants. Thus, the first phase of the experiment was an exemplar generation task and the second phase was a classification task.

Participants were then given a second set of cards with a different structure and repeated this procedure.

4.4. Results and discussion

When discussing the model predictions and human data, we will refer to novel combinations of observed parts, corresponding to the white cells in Fig. 5a, as valid exemplars. We will refer to all other responses as invalid exemplars. This section presents the results of the generation task. The results of the rating task are included in the section following Experiment 2 that discusses the copy-and-tweak account.

4.4.1. Structured sampling model predictions

The predictions of the structured sampling model are shown in the first plot of Fig. 6b. The model predicts that the eight most probable responses are the eight valid exemplars. Because the training set is small and the task is relatively unconstrained, the model also predicts a relatively large proportion of invalid exemplars, indicated by the white bars. However, the model assigns no single invalid exemplar more than a negligible probability ($\sim 3 \times 10^{-4}$). All of the plots in Fig. 6 include two bars for invalid exemplars labeled “Invalid (C)” and “Invalid (R),” corresponding to invalid exemplars that are at least as common as or rarer than the least common valid exemplar.

4.4.2. Human responses

The human responses for each condition are shown in Fig. 6a. As predicted by the structured sampling model, the majority of responses in each condition were the valid exemplars. Moreover, these exemplars were generated with approximately equal frequency in all conditions, with no clearly favored exemplar.

When viewing these results, it is important to realize that the labels corresponding to the rows and columns in Fig. 5a are arbitrary and therefore exemplars cannot be directly compared between conditions. For example, any one of the valid exemplars in a condition could have been designated as exemplar (1,3). Because each condition used different sets of letters, it does not make sense to

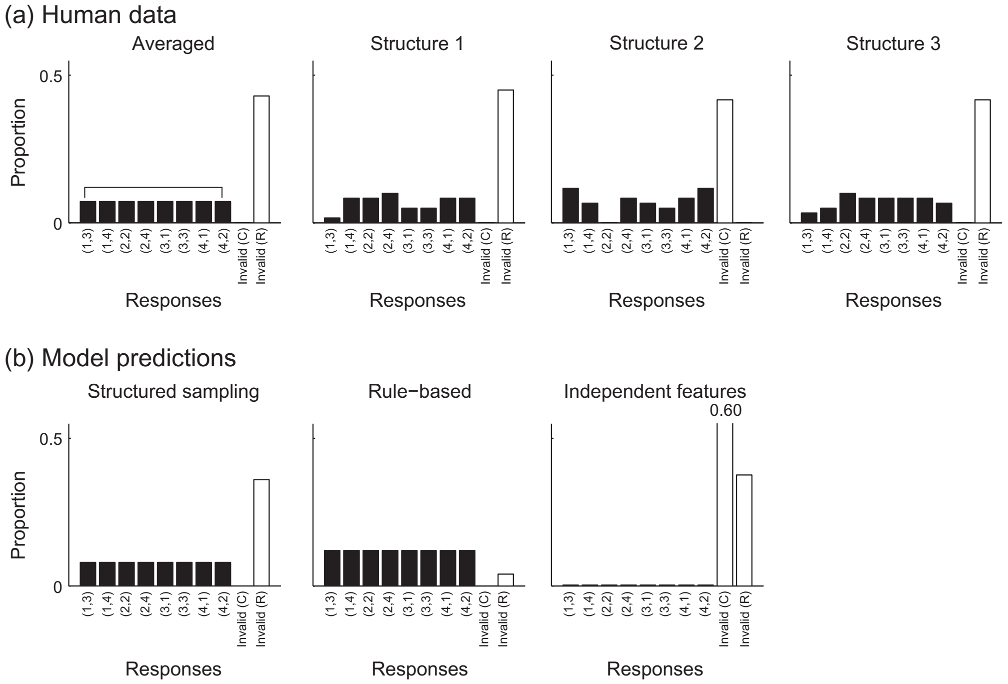

Fig. 6. Experiment 1 results and model predictions. (a) Human responses, averaged and for each of the three conditions. (b) Model predictions for the structured sampling model and the two sampling-based comparison models. The bar labeled "Invalid (C)" is for invalid exemplars that were generated more frequently than the least common valid exemplar (i.e., "common" invalid exemplars). The bar labeled "Invalid (R)" is for invalid exemplars that were generated less frequently than the least common valid exemplar (i.e., "rare" invalid exemplars). In the independent features plot, the Invalid (C) bar has a value of 0.60 but has been truncated in order to maintain the same vertical scale across all plots.

compare individual exemplars across conditions. The results can be combined, however, by averaging over all possible ways of mapping one condition onto another, as shown in the first plot of Fig. 6a. In the current experiment, every white cell in Fig. 5a corresponds to an exemplar made from two parts that were each observed twice. Thus, exemplar (1,3) from one condition could be mapped onto any exemplar from another condition, leading to the uniform appearance of the averaged results. The bracket in the plot highlights the fact that these exemplars cannot be distinguished in the averaged results.

Overall, the majority of responses (57%) were valid exemplars. Consistent with the structured sampling model, the eight most common responses were the eight valid exemplars. Of the invalid exemplars that people generated, only one was generated twice and no other exemplar was generated more than once. This result suggests that participants did not consistently use any alternative generation strategy that is not captured by the structured sampling model. Although no single invalid exemplar was consistently generated, the number of invalid exemplars generated was relatively high. This result is again consistent with the structured sampling model, which predicts a relatively large number of rare invalid exemplars (see the Invalid (R) bars in Fig. 6).

4.4.3. Sampling-based comparison model predictions

Predictions for the rule-based and independent features models are shown in Fig. 6b. Because the training data in this experiment used uniform distributions of parts, the rule-based model performs reasonably well, consistent with the intuition that inferences about graded slot distributions are not necessary in order to generate valid novel exemplars in this task. Just like the structured sampling model, the rule-based model correctly predicts the eight most common human responses. However, the rule-based model generates more confident predictions than the structured sampling model. The model's confidence is relatively high because it does not consider alternative parts that were not observed in the training data.

The independent features model, however, shows considerable uncertainty in its predictions and does not appear to capture the constraints that guide human inferences. In particular, as indicated by the Invalid (C) bar, exemplars that were generated more often than the eight valid exemplars accounted for 60% of the model's responses. Because the independent features model does not account for our data, we can conclude that human responses to the exemplar generation task depend critically on discovering the underlying structure of the category.

4.4.4. Graphics tablet results

Data from participants who illustrated their responses using graphics tablets allowed us to examine the method by which people constructed their responses. As described earlier, the sample-by-parts account predicts that participants will generate novel exemplars by sampling the two parts of each exemplar from independent slot distributions. The account therefore suggests that participants will tend to write both letters of one part before writing the letters of the second part. We might also expect that the pause between writing the first and second letters of the same part would tend to be shorter than the pause between writing the second letter of the first part and writing the first letter of the second part. Note that participants were not constrained in any way when illustrating their responses, and that valid exemplars could be generated by writing the letters in any order and at any time.

We coded a response as consistent with the letter order prediction if the first two letters written formed a complete part. Of 114 responses from 19 participants who used a graphics tablet, 88 (77%) were consistent with the model's sampling procedure, well above the chance level of 33% for our coding scheme, Binomial test $p < .001$.

To evaluate the timing prediction, we recorded the times at which participants began writing the four letters for each of their responses. Computing the differences between adjacent times produces three time intervals for each response. If participants paused between the parts, then the second interval (i.e., the interval between writing the second and third letters) should be longer than the first and third intervals. Because this prediction applies only to responses that were drawn one part at a time, we restricted the analysis to the 88 responses that met this condition. Of these responses, the second interval was the longest of the three intervals in 71 (81%) cases, well above the chance level of 33%, Binomial test p <.001. On average, participants paused an additional 8.9 s (SD = 20.8) between parts than they did within parts.

5. Experiment 2: probabilistic exemplar generation

The results for Experiment 1 suggest that people recognized that the training exemplars were composed of discrete parts, and were able to generate novel exemplars by combining parts that had not been previously paired. The experiment therefore supports our hypothesis that exemplar generation can be achieved by combining familiar parts, but does not directly address whether novel exemplars are generated by sampling from probability distributions over parts. Experiment 2 evaluates the probabilistic aspect of the sampling account. In this experiment, we replaced the balanced set of frequencies used in Experiment 1 with a skewed set shown in Fig. 5b. If people generate new exemplars by sampling from probability distributions over parts, then the part distributions across the novel exemplars should match the skewed parts distribution across the training exemplars.

5.1. Method

The design and procedure in Experiment 2 were identical to Experiment 1. The two experiments differed only in the set of exemplars shown to participants. In the training set for Experiment 2, one part in each slot appeared three times, two parts in each slot appeared two times, and one part in each slot appeared once, as shown in Fig. 5b. Thirty Carnegie Mellon University undergraduates completed the experiment for course credit.

5.2. Results and discussion

5.2.1. Structured Sampling Model Predictions

The predictions of the structured sampling model are shown in the first plot of Fig. 7b. The model once again predicts that the eight most probable responses are the valid exemplars, corresponding to the white cells in Fig. 5b. Because the part frequencies were varied in Experiment 2, the model predicts different relative probabilities for the eight valid exemplars. For instance, exemplar (1,1) is predicted to be most probable because it is composed of the two most frequently observed parts in the training set.

As for Experiment 1, the model predicts a relatively large proportion of invalid responses, although no individual invalid response is predicted to have more than a negligible probability ($\sim 4 \times 10^{-4}$).

5.2.2. Human responses

The human responses for each condition are shown in Fig. 7a. The majority of responses in all conditions were the valid exemplars, corresponding to the white cells in Fig. 5b. The most common response in all conditions was exemplar (1,1), followed by exemplars (2,2) and (3,3), as predicted by the structured sampling model.

As with the results in Experiment 1, not all exemplars can be directly compared between conditions. Unlike Experiment 1, however, some exemplars can be aligned across conditions. For example, exemplar (1,1) is the only exemplar made of two parts that each appeared three times in the training set. Exemplars (2,2) and (3,3), however, are made of parts that were each observed twice, and there-

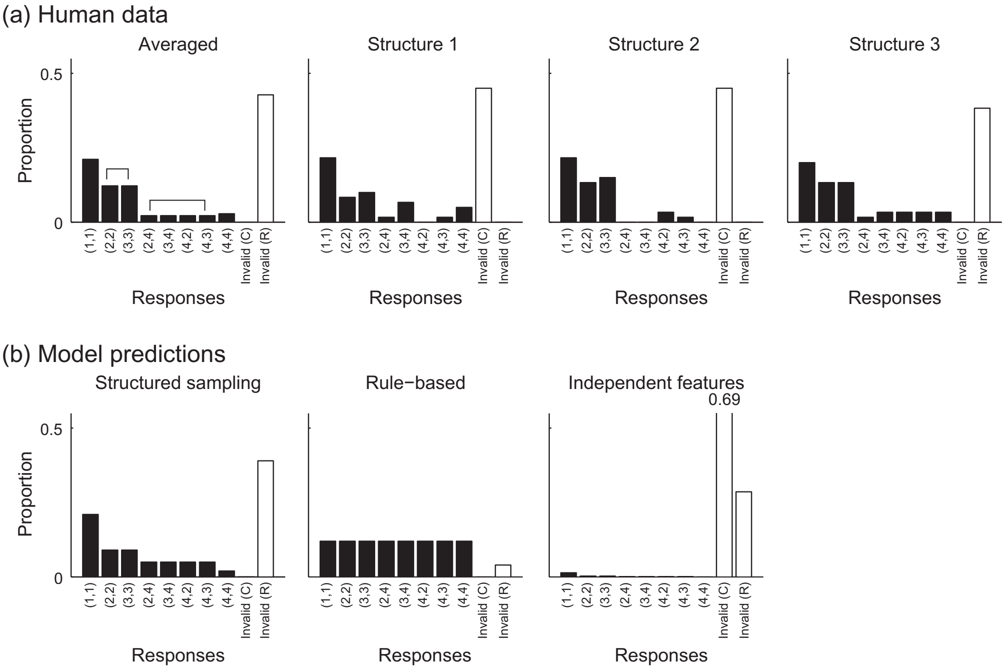

Fig. 7. Experiment 2 results and model predictions. (a) Human responses, averaged and for each the three conditions. (b) Model predictions for the structured sampling model and the two sampling-based comparison models.

fore must be averaged together to compare different conditions. These averaged groups are indicated by brackets in the combined results shown in the first plot of Fig. 7a.

Overall, the majority of responses (57%) were the valid exemplars. These exemplars were also the eight most common responses, as predicted by the structured sampling model. Moreover, the relative frequencies of valid exemplars in the human responses match the predictions of the structured sampling model, with the exception of exemplar (4,4), which was generated by participants about as often as other low-probability valid exemplars. Of the invalid exemplars that people generated, only one was generated twice and no other exemplar was generated more than once. Thus, just as in Experiment 1, participants do not appear to have consistently used any alternative exemplar generation strategies that do not depend on the category structure.

5.2.3. Sampling-based comparison model predictions

The results from Experiment 1 were approximately consistent with both the structured sampling model and the rule-based model. Therefore, the critical comparison in Experiment 2 was between these two models. Predictions of the rule-based model are shown in the second plot of Fig. 7b. Because the rule-based model does not take frequency information into account, it makes identical predictions for Experiments 1 and 2. Frequency information was not central to the category in Experiment 1, which explains why the rule-based model provided a reasonable account of the data in that experiment. Experiment 2, however, suggests that people take feature frequencies into account when generating novel category exemplars, a result that is predicted by the structured sampling model but not the rule-based model.

For completeness, we generated predictions for the independent features model, even though this model performed poorly in Experiment 1. The independent features model focuses *only* on feature frequencies and therefore might be expected to fare better in the current experiment. The model's predictions are shown in the third plot of Fig. 7b. Just as in Experiment 1, the model shows considerable uncertainty in its predictions, although the model does correctly predict different relative probabilities for the top human responses. These probabilities, however, are extremely small, which is why the model's predictions are difficult to detect in the figure. The model predicts that exemplar (1,1) will be the most common response, but incorrectly predicts that many invalid exemplars will be generated more frequently than the remaining seven valid exemplars (see the Invalid (C) bar). These invalid exemplars are composed of individual features from the most frequent parts but are not consistent with the category structure. This failed prediction, as well as the model's general failure to capture the level of certainty in people's responses, strongly suggests that people take the structure of the category into account when generating novel exemplars.

5.2.4. Graphics tablet results

Data from participants who illustrated their responses using graphics tablets included 132 responses from 22 participants. The same analysis used for Experiment 1 revealed that 116 (88%) of these responses were drawn one part at a time, Binomial test $p < .001$. Of the 116 responses drawn one part at a time, 80 (69%) were drawn with the longest pause between letters occurring between parts, Binomial test $p <.001$. On average, the pause between drawing successive letters was 9.2 s longer ($SD = 17.8$) between parts than within parts. These results reinforce the conclusion from Experiment 1 that people generated novel exemplars in a way consistent with the sample-by-parts account. In addition, Experiment 2 shows that our previous conclusion generalizes to cases in which the distributions over parts are not uniform.

5.2.5. Individual differences

The structured sampling model is meant to describe aggregate human behavior, which is why all of our analyses so far have focused on averaged results. However, these analyses might have obscured individual differences in the data. To address this issue, we recorded the number of valid exemplars, out of three, that each participant generated. Histograms of these counts for Experiments 1 and 2 (collapsed across conditions) are shown in Fig. 8. Nearly all of participants' responses in Experiment 1 can be classified into two groups: all valid exemplars or all invalid exemplars. This is true to a lesser extent for Experiment 2, in which a majority of responses fall into these two groups. These results suggest

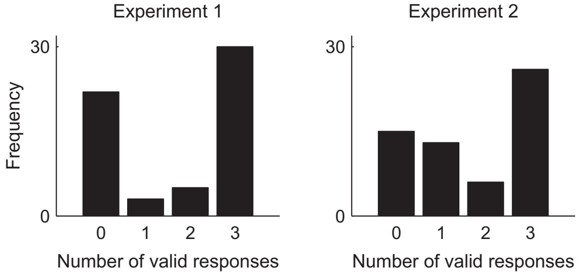

Fig. 8. Frequency of participants' valid responses count for Experiments 1 and 2.

that there was a relatively large minority of participants (those that generated no valid exemplars) that either failed to learn the category structure or deliberately disregarded it.

The structured sampling model does not account for the responses of individuals who generated no valid exemplars, but it is likely that the participants in this category relied on many different idiosyncratic strategies. As noted previously, no single invalid exemplar was generated more than two times in either experiment, strongly suggesting that participants did not detect any regularities in the category that are not captured by the model. Recall also that the task was relatively unconstrained. Participants were not told that they should only use letters they had seen before, or even that their responses should be arranged in a circle. The space of all possible sets of four letters found in the training set is large ($16^4 \approx 65{,}000$ possible exemplars), but a small number of participants used letters that were not in the training set, suggesting that they were considering an even larger hypothesis space. Given the vast number of logically possible responses, it is striking that the majority of responses were so consistent.

5.3. Discriminative accounts of exemplar generation

We now consider the extent to which our results from Experiments 1 and 2 can be explained by the discriminative accounts introduced earlier.

5.3.1. The randomly-sample-and-score account

As described earlier, the randomly-sample-and-score (RSS) account proposes that any discriminative model of classification can be converted into a model of generation by randomly sampling exemplars from the full feature space and then scoring them using the discriminative model's classification function. These scores can then be used to decide which of the randomly sampled exemplars to generate. When implementing the RSS model in this section, we used the classification distribution induced by the structured sampling model to score the sampled exemplars. Other classification functions could be used, but our choice of classification function enables a direct comparison between the RSS model and the structured sampling model: both models learn the underlying category in the same way, and differ only in how they generate exemplars.

Here we consider one RSS model that repeatedly draws one sample, scores it, and makes a decision about whether to generate it before sampling another. Section A.2.5 of the Appendix considers an additional RSS model that uses a different sampling approach. One way to decide whether to generate a sampled item is to set a fixed acceptance threshold and generate all samples that score above this threshold. This approach could perform well in Experiment 1 with an appropriately chosen threshold, but would be unable to account for the frequency effects in Experiment 2 because all exemplars exceeding the threshold would be generated equally often. We therefore assumed that the probability of generating a sampled exemplar increases as its score increases. We also introduced a parameter that specifies a lower bound on the probability of generating each exemplar. This parameter can be interpreted as a parameter that specifies the expected number of samples $E(N)$ required to generate three exemplars.

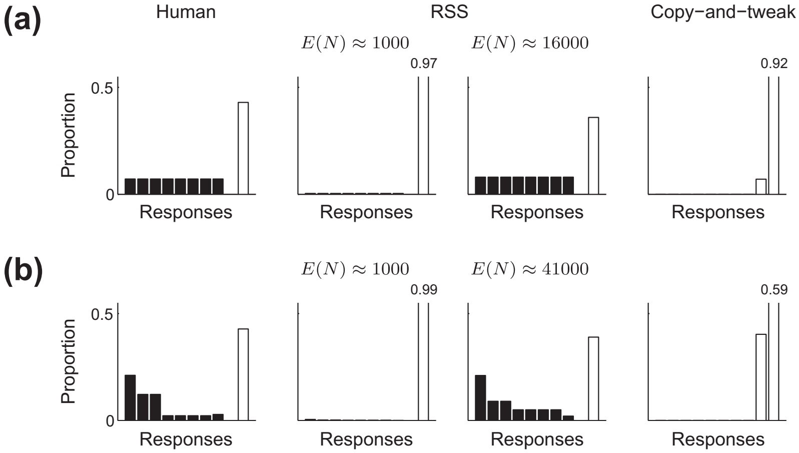

Fig. 9. Human data and predictions for the RSS and copy-and-tweak models for (a) Experiment 1 and (b) Experiment 2. The RSS model makes different predictions depending on the expected number of samples, $E(N)$, needed to generate three exemplars. RSS model predictions are shown for two different values of $E(N)$. The second set of RSS model predictions is based on the maximum possible value of $E(N)$, which applies if the probability of generating a sample is directly proportional to its score.

The model's predictions for Experiments 1 and 2 are shown in Fig. 9. The figures show predictions for two values of the sample size parameter.[3] When the probability of generating a sampled item is directly proportional to its score, the RSS model is provably equivalent to the structured sampling model. This result holds in Experiment 1 when the sample size parameter is around 16,000 on average, and in Experiment 2 when the sample size parameter is around 41,000 on average. Fig. 9 confirms that the RSS and structured sampling models generate identical predictions when the sample size parameter is set sufficiently high. Although the models generate identical predictions, the two are different in one important respect. The RSS model must draw thousands of samples from a uniform distribution over the space of exemplars, but the structured sampling model draws exactly three samples because it samples directly from the generation distribution.

Any model that draws thousands of samples seems psychologically implausible, and we therefore explored the predictions of the RSS model when the sample size parameter is smaller. Fig. 9 shows that the model fails to account for human responses when $E(N) \approx 1000$. In this case, the model under-generates the eight valid exemplars that were preferred by humans in both experiments. A sample size of 1000 is arguably still too large to be psychologically plausible, but reducing the sample size further would only accentuate the difference between model predictions and human judgments. We can therefore conclude that the RSS model does not account for our data when the sample size parameter is set to a psychologically plausible value.

5.3.2. The copy-and-tweak account

As described earlier, the copy-and-tweak approach converts an exemplar model of classification into a model of generation. The approach proposes that new exemplars are generated by copying one of the exemplars stored in memory and then modifying some features of the copy. When we introduced the category used for Experiments 1 and 2, we noted that our stimuli were carefully chosen to distinguish between the structured sampling account and the copy-and-tweak account. The structured sampling model is capable of learning that pairs of letters are grouped into cohesive parts and therefore predicts that people will tend to generate novel combinations of parts. In contrast, a copy-and-tweak model will generate many exemplars that are not valid exemplars. For example, after observing the four exemplars in Fig. 4b, a copy-and-tweak model might produce a new exemplar by copying the first exemplar and changing the top letter from D to X.

[3] These values are approximate, but exact values are provided in the Section A.2.4 of the Appendix.

We implemented a copy-and-tweak model based largely on the GCM (Nosofsky, 1986). The GCM makes classification judgments by computing the similarity between a novel exemplar and all of the stored exemplars, and similarity is computed as a function of pairwise distance between exemplars. Because our stimuli used discrete non-ordinal features, we used the Hamming distance, which measures the number of features that are different. As explained earlier, the GCM is equivalent to a generative model that maintains a probability distribution with peaks at the observed exemplars (Ashby & Alfonso-Reese, 1995; Jäkel et al., 2008). The generative model that corresponds to the GCM maintains a probability distribution over a continuous similarity space, and we implemented an analogous approach that maintains a probability distribution over our space of discrete flu genomes.

The predictions of this copy-and-tweak model are shown in the last column of Fig. 9. For both experiments, the model under-generates the eight valid exemplars. In addition, the predictions for Experiment 2 show that the model generates many invalid exemplars more frequently than the valid exemplars.

In addition to failing to account for the results of the generation task, the copy-and-tweak model also fails to account for the results of the classification task. Model predictions for this task are shown in the second and third columns of Fig. 10. These predictions were generated by computing the probability of category membership (structured sampling and RSS models), or the sum of similarities (copy-and-tweak model), for every item in the rating task. These values were then treated as ratings and converted to z-scores to make them comparable to the human data, described below. Fig. 10

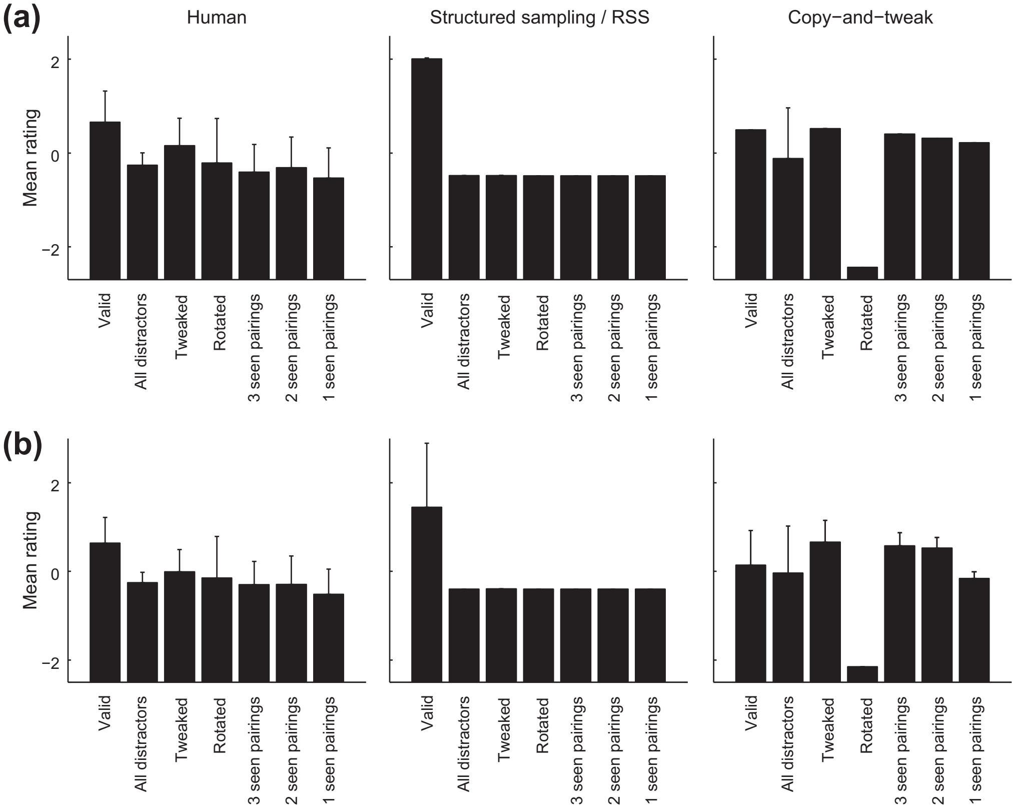

Fig. 10. Mean ratings and model predictions (converted to *z*-scores) for the rating task for (a) Experiment 1 and (b) Experiment 2. All error bars denote one standard deviation.

shows predictions for both valid exemplars and some specific types of invalid exemplars (distractors). These distractors included exemplars that were created by changing a single feature from a training exemplar (tweaked exemplars), exemplars that were generated by rotating a valid exemplar 90° to violate the category structure, and exemplars that included between one and three previously seen pairs of features. In both experiments, the structured sampling and RSS models make a straightforward prediction, assigning high ratings to valid exemplars and uniformly low ratings to distractors. Note that these models make identical predictions because they rely on the same classification distribution. In contrast, the copy-and-tweak model assigns approximately equal ratings to valid exemplars and all but the rotated invalid exemplars.

The human data are shown in the first column of Fig. 10. Because different participants used the rating scale differently, we converted each participant's responses to z-scores. The figure displays the means of these scores for each type of rating item. Four set of ratings from Experiment 1 and two sets from Experiment 2 were removed because the participants failed to rate every item. In both experiments, there was a significant difference between the mean scores per participant for valid (Experiment 1: $M = 0.66$, $SD = 0.66$; Experiment 2: $M = 0.64$, $SD = 0.58$) and invalid (Experiment 1: $M = -0.26$, $SD = 0.27$; Experiment 2: $M = -0.25$, $SD = 0.23$) exemplars (Experiment 1: $t(55) = 7.40$, $p < .001$; Experiment 2: $t(57) = 8.33$, $p < .001$). Of particular interest are the ratings for the tweaked exemplars and the exemplars with three previously seen feature pairings. If participants based their judgments only on feature similarity to the training exemplars without regard to the category structure, they would give high ratings to these exemplars. In both experiments, however, participants gave significantly higher ratings to the valid exemplars than to both the tweaked exemplars (Experiment 1: $M = 0.16$, $SD = 0.59$, $t(55) = 4.12$, $p < .001$; Experiment 2: $M = -0.01$, $SD = 0.50$, $t(57) = 6.17$, $p < .001$) and the exemplars with three previously seen feature pairings (Experiment 1: $M = -0.41$, $SD = 0.60$, $t(55) = 7.71$, $p < .001$; Experiment 2: $M = -0.30$, $SD = 0.52$, $t(57) = 7.17$, $p < .001$).

Overall, the classification results are broadly consistent with the structured sampling and RSS models but not the copy-and-tweak model. Our data therefore suggest that the copy-and-tweak model fails to account for both the generation data and the classification data. Because this model is a simple extension of the exemplar-based approach to categorization, our results suggest that exemplar models are unlikely to provide a complete account of people's knowledge about categories. The main reason why the copy-and-tweak model fails to account for our data is that it does not capture the compositional structure of the category used in our experiments. People may indeed store specific exemplars, as proposed by the exemplar-based approach, but in addition, they appear to acquire a category representation which specifies that these exemplars are constructed out of coherent parts.

By departing from the representational assumptions of exemplar models, one could develop a more sophisticated copy-and-tweak approach that accounts for our data. Consider, for example, a model that learns that exemplars are built out of parts, and that uses tweaks operating at the level of parts (e.g., exchanging one part for another) rather than at the level of individual features. A model of this kind could generate new exemplars by copying a previously observed exemplar, randomly choosing one or more parts, and replacing these parts by sampling from the empirical distribution over parts. The resulting model, however, would rely on the same basic assumptions as the sample-by-parts approach, and could therefore be viewed as an instantiation of this approach.

6. Experiment 3: category generation

The previous two experiments suggested that the sampling account helps to explain how people generate novel exemplars of known categories. We now move one level up in the hierarchy in Fig. 1 and explore category generation in two experiments that ask participants to generate exemplars of novel categories. As before, our presentation of Experiments 3 and 4 will be organized around the sampling account of generation. After presenting Experiment 4, we will discuss the extent to which the RSS and copy-and-tweak approaches can account for both experiments.

The two main goals of Experiment 3 are as follows. First, we illustrate how the sampling account can be applied to a specific category generation task by developing a model that learns category constraints and then generates new categories that respect these constraints. Second, we evaluate this model by comparing its predictions to human responses. Earlier we reviewed past research on learning category constraints. Experiment 3 asks whether people respect these constraints when generating novel categories. To explore this question, we designed sets of categories where the same two dimensions were correlated within each category. The sampling-based model described later predicts that people will generate new categories that preserve the correlations present in the training categories. To test this prediction, we varied the sign of these correlations and explored whether people preserved these correlations in the novel categories that they generated.

6.1. A category generation task

The exemplar generation task in the previous experiments used a category with discrete feature values. Our computational account, however, can handle both discrete and continuous features. To illustrate this flexibility, our second set of experiments considers category exemplars with continuous features. The categories were sets of geometric figures that we called crystals, some examples of which are shown in Fig. 11. In Experiment 3, a learner observes two categories of these crystals (the training categories), and is then asked to generate a novel third category of crystals.

We varied three dimensions of the crystals: hue, length, and saturation. In all training sets, the hue of the crystals was held constant within-category but varied between-category. The relationship between the length and saturation dimensions was varied to produce training sets in which these two dimensions were positively correlated (Fig. 11a), negatively correlated (Fig. 11b), or uncorrelated (Fig. 11c). Thus, although categories in a given training set had different hues, all of the categories exhibited the same relationship between length and saturation. A learner could therefore generate a new category by selecting a single hue and then varying the length and saturation values of the category's exemplars to produce the appropriate correlation.

6.2. The hierarchical sampling model

We now sketch how the sampling account can be applied to the category generation task using the crystals just described. Section A.3.1 of the Appendix contains a full specification of the model.

We begin by formulating the task using our computational notation. Each category y corresponds to a distribution over the three dimensions, the exemplars and category labels $(\mathbf{x},\mathbf{y})$ represent the training data, and the learner must generate a set of category exemplars $\tilde{\mathbf{x}}$ from a novel category $\tilde{y}$. The sampling account proposes that these exemplars are generated by first sampling from the category generation distribution $p(\boldsymbol{\theta}_{\tilde{y}}|\boldsymbol{\phi})$ to generate a new category, then sampling several exemplars from the resulting exemplar generation distribution, $p(\tilde{x}|\boldsymbol{\theta}_{\tilde{y}})$. Recall that $\boldsymbol{\phi}$ denotes the higher-order category parameters and $\boldsymbol{\theta}_{\tilde{y}}$ denotes the parameters of the new category. In order to generate a new category, $\boldsymbol{\phi}$ must be inferred.

We assume that each category y can be described by a three-dimensional normal distribution with a mean μ_y and a covariance matrix Σ_y. The parameter vector $\boldsymbol{\theta}_y$ for each category therefore corresponds to the pair (μ_y, Σ_y). We assume that each $\boldsymbol{\theta}_y$ was generated by starting with some initial

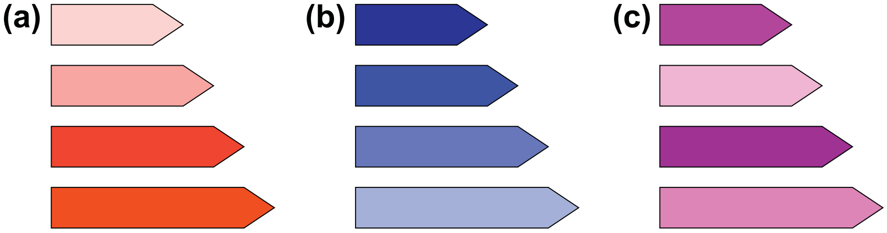

Fig. 11. Example stimuli for the category generation task used in Experiments 3 and 4. The categories were described as sets of crystals. Hue was held constant within-category but was varied between-category. The length and saturation dimensions were either (a) positively correlated, (b) negatively correlated, or (c) uncorrelated.

"template" $\boldsymbol{\phi} = (\mu_0, \Sigma_0)$ that defines the higher-order category, and randomly tweaking some of its values. This idea is illustrated in Fig. 12. In the figure, the Σs are depicted in matrix form. The training categories are assumed to have been generated by randomly tweaking the mean μ_0 and each cell of the covariance matrix Σ_0 (subject to some basic constraints specified in the Appendix) to produce a category mean μ_y and covariance matrix Σ_y. The exemplars $\mathbf{x}$ are then assumed to have been generated by sampling from normal distributions with these category parameters. The Appendix shows how these assumptions can be used to infer the category parameters $\boldsymbol{\theta}_y$ and the higher-order category parameters $\boldsymbol{\phi}$. This model is an instance of a hierarchical Bayesian model, a type of model that has previously been used to model category constraint learning (Perfors & Tenenbaum, 2009; Kemp, Perfors, & Tenenbaum, 2007) but not category generation.

After inferring the category and higher-order category parameters, new categories can be generated by applying the same generative assumptions required for inference: we randomly tweak the values in $\boldsymbol{\phi}$ to produce a set of novel category parameters $\boldsymbol{\theta}_{\tilde{y}}$, then generate some exemplars $\tilde{\mathbf{x}}$ of the novel category by sampling from a normal distribution with the novel category parameters. This procedure is shown visually in Fig. 12. Because this model represents a hierarchy of categories, we will refer to it as the hierarchical sampling model.

The sample-by-parts account that we proposed earlier does not apply in this setting because the crystal categories do not have the same kind of modular structure as the genome categories in Experiments 1 and 2. Later we discuss alternative mechanistic approaches that could be used to sample from the distributions specified by the hierarchical sampling model. Here, however, we evaluate the hierarchical sampling model by comparing its predictions to the predicitions of an alternative sampling-based model described in the next section.

6.2.1. Sampling-Based Comparison Model

The hierarchical sampling model relies critically on the category hierarchy shown in Fig. 12. To explore whether humans make use of this hierarchy when generating new categories, we will evaluate an alternative approach that does not rely on a hierarchy of categories but is otherwise identical to the hierarchical sampling model.

The independent categories model. The independent categories model does not perform inference at the level of higher-order category parameters $\boldsymbol{\phi}$. Instead, the model assumes that the parameters $\boldsymbol{\theta}_y$ for each category are drawn from a distribution that does not change over the course of the experiment. As a result, the model is capable of generating new categories and exemplars, but these categories will not be constrained by the training data.

The independent categories model is closely related to a Gaussian mixture model. Gaussian mixture models are standard statistical tools for clustering data drawn from different populations and have previously been used to model human category learning (Rosseel, 2002; Vallabha, McClelland, Pons, Werker, & Amano, 2007). Although Gaussian mixture models are useful for some purposes, we predict that the independent categories model will not account for our category generation data as well as the hierarchical sampling model. If the independent categories model provides a better

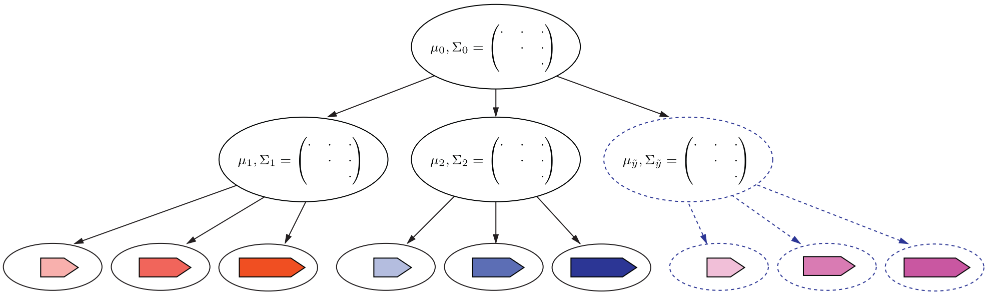

Fig. 12. The hierarchical sampling model used for our category generation task.

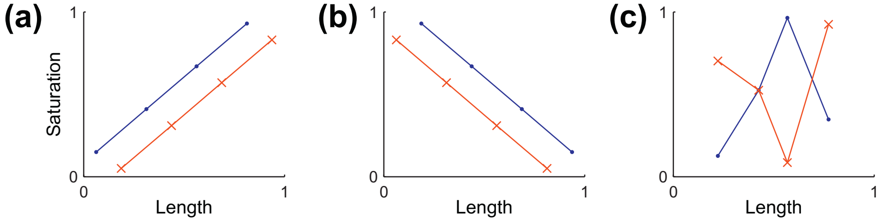

Fig. 13. Stimuli used in Experiment 3. The plots show the length and saturation values of the exemplars for the training categories in conditions (a) r_+, (b) r_-, and (c) r_0. In each plot, the exemplars of one category are represented by crosses and the exemplars of the other category are represented by dots. The exemplars of each category are connected by a line for clarity.

fit to our data, this result will suggest that people have not learned the category constraints, or that they have not applied them as predicted by our computational account.

6.3. Method

6.3.1. Participants

Twenty-two Carnegie Mellon University undergraduates completed the experiment for course credit.

6.3.2. Design

The experiment included three conditions that used three different training sets: one in which the length and saturation dimensions of the training categories were positively correlated, one in which they were negatively correlated, and one in which they were uncorrelated. We will refer to the three conditions as r_+, r_-, and r_0, respectively. Each participant completed all three conditions in random order.

6.3.3. Materials

The entire experiment was conducted using a graphical interface on a computer. In each condition, participants were presented with two training categories, each consisting of four exemplars. Each dimension varied on a scale from 0 to 1. For the length dimension, 0 corresponded to the shortest allowable length of the crystals and 1 corresponded to the longest allowable length. For saturation, 0 corresponded to the minimum allowable saturation, which was restricted to be slightly above a true 0 value (which would have appeared white), and 1 corresponded to full saturation. For the hue dimension, both 0 and 1 corresponded to red, as the hue scale is naturally circular. The hue and saturation dimensions corresponded to the hue and saturation coordinates of the standard HSV (hue-saturation-value) color coordinate system. Proximity in HSV space does not always align with perceptual similarity, but HSV space is adequate for an experiment that focuses on correlations between dimensions rather than precise distances in color space.

Fig. 13 shows the training stimuli used for the three conditions. As shown in the figure, values along the length and saturation dimensions were perfectly correlated in the r_+ and r_- conditions. The corresponding correlation for the r_0 condition was close to zero but weakly positive.[4] The hue dimension for each training category was chosen at random from a set of six distinct hues, with no hue repeated during each run of the experiment. The order of exemplars and the order of the training categories was randomized.

[4] The stimuli in this condition were chosen to provide a reasonable amount of variation along both the length and saturation dimensions but to have no feature value repeated within each category. Because the number of exemplars was small, these constraints ruled out all configurations where the correlation between the length and saturation dimensions was exactly zero. The two categories had correlations of 0.37 and 0.12.

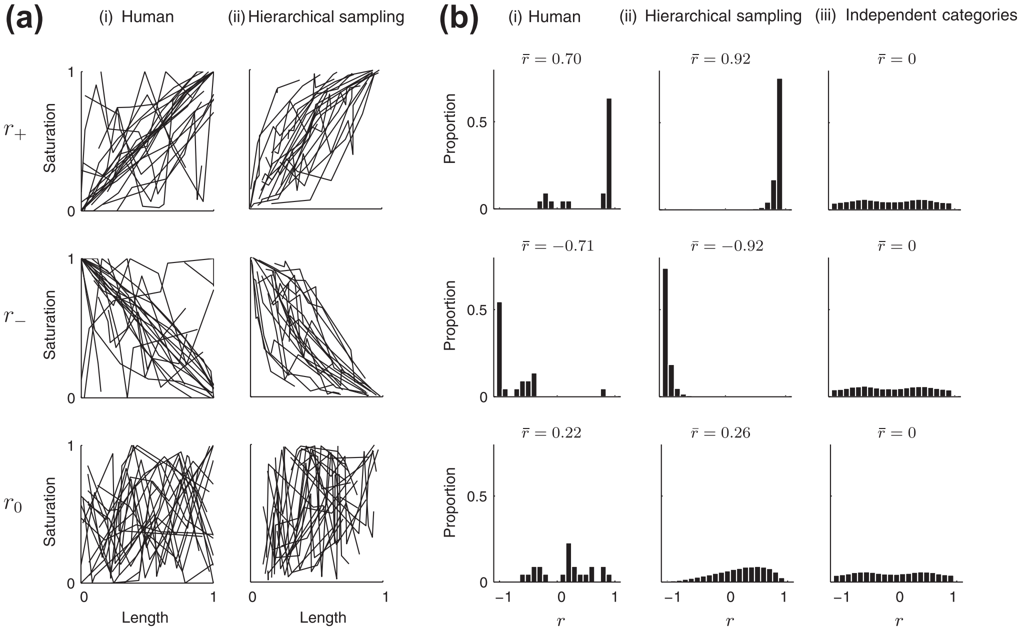

Fig. 14. Human data and model predictions for length-saturation correlations in Experiment 3. (a) Experiment results and predictions of the hierarchical sampling model. (i) The full set of participant-generated crystal categories for each condition. Each line in the plots corresponds to a category of six generated crystals. (ii) Results from a representative model run in which one category was generated for each participant. See the text for details about this simulation. (b) A comparison between aggregate human data and predictions of the hierarchical sampling and independent categories models. (i) Histograms of correlation coefficients for the human data in each condition. (ii) Predictions of the hierarchical sampling model. (iii) Predictions of the independent categories model.

6.3.4. Procedure

Participants were told that a space probe was sent to each of several planets to collect crystals. The two training categories were given arbitrary alphanumeric labels, and participants were told that the crystals in these categories were samples of two different types of crystal from the same planet. The screen was partitioned into three columns separated by vertical lines. The first training category was confined to the first column and the second training category was confined to the second column. Participants saw one training category and were allowed to examine it as long as they wished before pushing a button to reveal the second training category. In both cases, the individual crystals appeared in a randomized order, but participants were allowed to organize and reorder them on the screen. When participants were ready, they pressed the button again to advance to the generation phase. Both of the training categories remained on the screen during the generation phase to reduce memory demands.

Participants were told that six samples of a third type of crystal had been collected from the same planet, and were asked to make a guess about what those samples looked like by generating their own crystals. Each training category had four crystals and we deliberately asked participants to generate a category with a different number of crystals (six) to prevent them from simply copying an entire existing category. Crystals were created by pressing a button that generated a random crystal. Its appearance could be adjusted by moving three independent sliders, corresponding to the three dimensions of length, hue, and saturation. Participants were allowed to create, adjust, or move their crystals in any order, and could edit previously created crystals before submitting their final responses.

Participants repeated this procedure two more times for each of the remaining two conditions. Each time, they were told that the crystals were collected by a different space probe sent to a different planet than before.

6.4. Results and discussion

6.4.1. Dimension correlations

The hierarchical sampling model is designed to learn relationships between dimensions that hold across categories. Predictions generated from the model are shown in Fig. 14b.ii, where the three rows correspond to the three conditions. The plots show posterior distributions on length-saturation correlations according to the model. See Section A.3.3 of the Appendix for details about how these predictions were generated. The model makes relatively strong inferences in conditions r_+ and r_- in favor of positive and negative correlations, respectively. In the r_0 condition, the model makes a much weaker inference, but still shows a bias toward positive correlations. This bias is consistent with the weakly positive correlation in the training data for this condition. For comparison, predictions of the independent categories model are shown in Fig. 14b.iii. Although this model is capable of generating new categories, those categories are not constrained in any way by the training data, and the model makes identical predictions for all three conditions.

The full set of participant-generated categories is shown in Fig. 14a.i. In the plots, each line traces one participant's six generated crystals. Our primary prediction was that participants' generated categories would preserve the dimensional correlations present in the two training categories. To evaluate this prediction, we computed the correlation between the length and saturation dimensions of each participant's generated category. The results of this analysis are shown in the histograms in Fig. 14b.i. As these panels show, most participants generated categories that preserved the dimensional correlations present in the training categories. Participants generated mostly positively correlated crystals in the r_+ condition ($\bar{r} = 0.70$), mostly negatively correlated crystals in the r_- condition ($\bar{r} = -0.71$), and crystals with more variable degrees of correlation, with a slight positive bias, in the r_0 condition ($\bar{r} = 0.22$). Wilcoxon signed-rank tests indicated that these sets of correlations were significantly different (p <.01 in all cases). The slight positive bias in the r_0 condition ($M = 0.22$, $SD = 0.44$) was also significantly greater than 0, $t(21) = 2.28$, $p < .05$, as predicted by the hierarchical sampling model. Overall, people behaved in a way consistent with the hierarchical sampling model, but not the independent categories model, suggesting that most people learned the category constraints and respected those constraints when generating novel categories.

For illustrative purposes, we used the hierarchical sampling model to generate one novel category of crystals for each of the training sets seen by the 22 participants in each condition of the experiment. The results of this simulation are shown in Fig. 14a.ii.[5] These results provide a qualitative snapshot of the model's predictions and suggest that the apparent variability of the human data in Fig. 14a.i is consistent with the model predictions in Fig. 14b.ii.

6.4.2. Dimension variability

Although the prediction of primary interest involves correlations between the length and saturation dimensions, the hierarchical sampling model also makes predictions about the variance along all three dimensions. Recall that the length and saturation dimensions varied considerably within each training category but that the hue was held constant. This property should seem highly non-accidental, and the hierarchical sampling model predicts that it will be replicated when generating a new category.

Fig. 15 shows model predictions of the standard deviations for the three category dimensions. Because neither the experimental results nor the model predictions differed significantly between conditions, the results shown in the figure were collapsed across all three conditions. Fig. 15a shows histograms of standard deviations for the participant-generated crystal categories. Wilcoxon signed-rank tests indicated that the set of standard deviations for the hue dimension ($M = 0.05$, $SD = 0.09$) was significantly different from the sets of standard deviations for the length dimension ($M = 0.32$, $SD = 0.07$), $Z = -6.95$, $p < .001$, and the saturation dimension ($M = 0.32$, $SD = 0.07$), $Z = -6.97$, $p < .001$. This result suggests that participants learned that the categories did not vary along the

[5] The results shown here are based on a single simulation, but are representative of several simulation runs that produced roughly equivalent results.

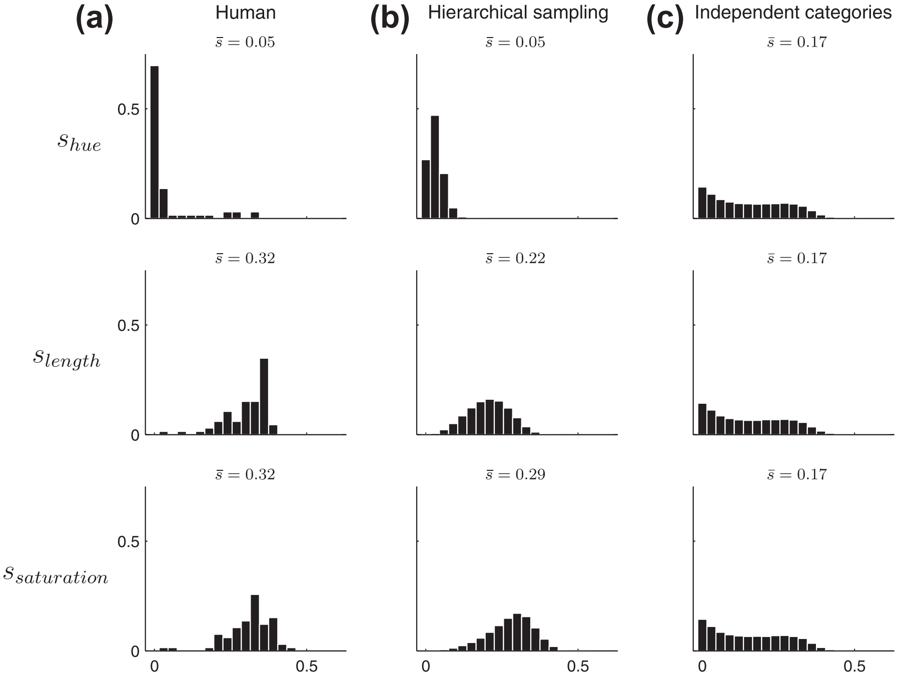

Fig. 15. Human data and model predictions for standard deviations in Experiment 3. (a) Histograms of standard deviations for each of the three category dimensions for participant-generated crystal categories. These results have been collapsed over all three experimental conditions. (b) Predictions of the hierarchical sampling model. (c) Predictions of the independent categories model.

hue dimension and applied this constraint when generating their own categories. The full set of results is consistent with the predictions of the hierarchical sampling model, shown in Fig. 15b. By contrast, the independent categories model cannot account for these results because it behaves identically in all three conditions, as shown in Fig. 15c. Because the model does not learn any category constraints, it does not change its diffuse prior distribution over covariance matrices.

7. Experiment 4: restricted category generation

The results for Experiment 3 suggest that people learned a relationship between dimensions that held across categories and preserved this relationship when generating new categories. The two main predictions of the hierarchical sampling model were supported: people generated categories that preserved the relationship between the length and hue dimensions that had been observed in the training categories, and they varied the hue dimension much less than the other two dimensions, again consistent with the training categories. Neither result was consistent with an alternative sampling-based model that treated categories as independent.

To deter participants from simply copying one of the observed categories, we asked them to generate a novel category with a different number of exemplars than were present in the two training categories. Experiment 3, however, did not completely rule out the possibility that participants generated a novel category by copying one of the observed categories and then duplicating some of the exemplars in this copied category. This approach would not explicitly recognize any dimensional correlations present in the observed categories, but would still preserve these correlations in the generated category. Experiment 4 was designed to rule out this alternative account of our data. The experiment was identical to Experiment 3 except that participants were required to generate their crystal categories in a novel region of the feature space, preventing participants from simply copying the observed categories. Even so, we predicted that participants would generate categories that preserved the correlations in the observed categories.

7.1. Method

Twenty-three Carnegie Mellon University undergraduates completed the experiment for course credit. The design and procedure were mostly identical to Experiment 3. In Experiment 4, however, a transformation was applied to the training categories that restricted them to one half (randomized across participants) of the length dimension, as shown in Fig. 16. In the generation phase of the experiment, participants were told that they would only be able to create crystals on the opposite side of the length dimension. This restriction meant that it was not possible to exactly copy any of the crystals in the training set. This experimental setup was inspired by a study on feature inference by Anderson and Fincham (1996). In one of their experiments, participants learned a correlation between two dimensions of a category and then made predictions about novel items drawn from a new region of the stimulus space.

7.2. Results and discussion

The results of primary interest are the correlations between the length and saturation dimensions. The full set of human data and model predictions is shown in Fig. 17. Panels a.i and a.ii compare the actual crystal categories generated by participants to an illustrative set of matched simulation samples (generated in the same way as for Experiment 3). Panels b.i and b.ii compare the correlations present in the human data to the predictions of the hierarchical sampling model. Although participants' inferences were not as consistent as in Experiment 3, most participants generated crystal categories that preserved the dimensional correlation present in the training set. As for Experiment 3, the overall set of results is broadly consistent with the hierarchical sampling model.

7.2.1. Toward a mechanistic account of category generation

Our data for Experiments 3 and 4 are consistent with the hierarchical sampling model, which proposes that people are able to sample from distributions over the exemplars of a novel category. So far, however, we have not discussed the mechanisms that allow people to sample from these distributions. This section provides some initial observations about the strategies participants used to generate categories in our experiments, and discusses the prospect of developing mechanistic models that capture these category generation strategies.

The computer interface used for Experiments 3 and 4 was not originally designed to record the complete sequence of actions participants took in order to generate a novel category. We therefore performed a follow-up experiment with 14 Carnegie Mellon University undergraduates that was identical to Experiment 3 from the perspective of a participant, but where the actions of each participant were recorded in full. The categories that these participants generated were consistent with the

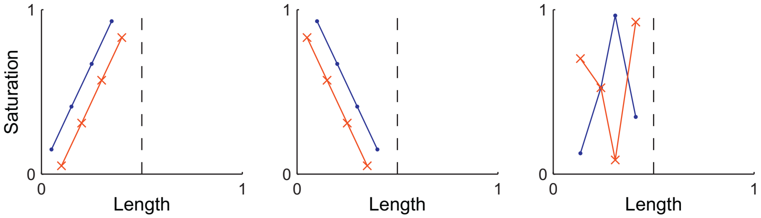

Fig. 16. Stimuli used in Experiment 4. In the training phase of the experiment, participants were shown crystals that were restricted to one half of the length dimension. In the category generation phase, crystals were restricted to the opposite half of the length dimension. As in Fig. 13, in each plot, the exemplars of one category are represented by crosses and the exemplars of the other category are represented by dots. The exemplars of each category are connected by a line for clarity.

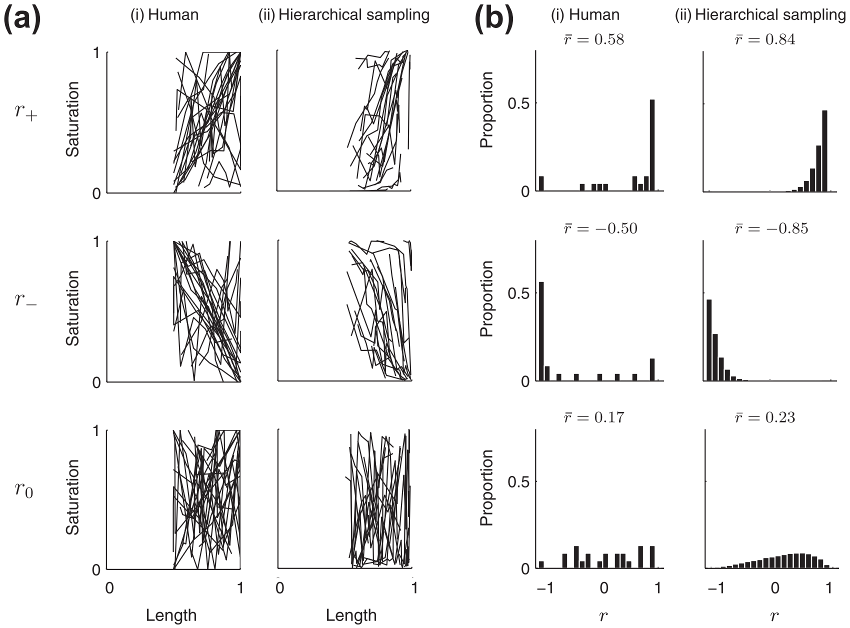

Fig. 17. Human data and model predictions for length-saturation correlations in Experiment 4. (a) Experiment results and predictions of the hierarchical sampling model. (i) The full set of participant-generated crystal categories for each condition. Each line in the plots corresponds to a category of six generated crystals. Participants were randomly restricted to one half of the length dimension, but in these plots some responses are shifted by 0.5 along the length dimension to make all responses comparable. (ii) Results from a representative model run in which one category was generated for each participant. (b) A comparison between aggregate human data and predictions of the hierarchical sampling model. (i) Histograms of correlation coefficients for the human data in each condition. (ii) Predictions of the hierarchical sampling model.

results reported for Experiment 3. We therefore focus here on the two most common strategies used to generate these categories.

Generating a novel crystal category requires a participant to specify 18 continuous values: three values along the dimensions of length, hue, and saturation for each of six crystals. The *dimension-focused* strategy considers the dimensions in sequence: for example, the length of the six crystals could be adjusted first, followed by the hue and the saturation. The *object-focused* strategy considers the crystals in sequence: for example, all three dimensions of the first crystal could be adjusted before moving on to the next crystal. We found that nearly all of the responses in the follow-up experiment were consistent with one of these two strategies.

The dimension-focused and object-focused strategies suggest a generalization of the sample-by-parts account, where the “parts” of a category are either the sets of feature values along different dimensions or the individual exemplars. Our data suggest that participants typically generated new categories either one dimension at a time or one object at a time. The hierarchical sampling model does not account for this result, and makes no prediction about the order in which participants generate the 18 values that characterize a novel category. This model, however, could be grounded in a mechanistic account that uses conditional sampling to capture the two generation strategies. For example, the dimension-focused strategy could be captured by a mechanistic account that samples the values along one dimension at random, then samples a value for the second dimension conditioned on the value for the first dimension, and finally samples a value for the third dimension conditioned on the values for the first two dimensions. Similarly, the object-focused strategy could be captured by a mechanistic account that samples an initial exemplar at random, then samples subsequent exemplars conditioned on the exemplars generated previously.

One final result emerging from Experiments 3 and 4 and the follow-up experiment is that many participants took care to generate categories that were "representative" of the relationship between dimensions. Often, for example, they chose length and saturation values that spanned these entire dimensions, which made any correlation between these dimensions especially apparent. In many cases, they also ordered the crystals on screen by increasing or decreasing length. The hierarchical sampling model does not predict these results: if six crystals are sampled independently from a novel category, the resulting set is not guaranteed to span the range of any of the three dimensions. Our results therefore suggest that a successful mechanistic account of category generation will need to go beyond the simplifying assumption that exemplars are generated independently, and to consider mechanisms that sample sets of exemplars in a way that is representative of the underlying category structure.

7.3. Discriminative accounts of category generation

We now return to the RSS and copy-and-tweak models and consider the extent to which these approaches can account for Experiments 3 and 4. We previously ruled out the RSS account of exemplar generation because it requires an inordinately large number of samples to account for our data. This problem is only magnified when moving from exemplar generation to category generation. In Experiment 3, for example, generating six crystals requires a learner to specify 18 continuous values. The number of perceptually distinguishable six-category samples is extraordinarily large, meaning that a RSS model that samples candidate categories at random is exceedingly unlikely to hit upon a suitable candidate. Therefore, the RSS account is not a viable alternative to the sampling account of category generation.

Our discussion of Experiments 1 and 2 showed that converting an exemplar-based model into a copy-and-tweak model could not account for our data. When applied to Experiments 3 and 4, the analogous approach would generate a new category by copying all of the exemplars in a previously observed category, then tweaking all of the copied exemplars. If necessary, exemplars could subsequently be added to the novel category by copying and tweaking an exemplar of that category. Experiment 4 was specifically designed to rule out this account. In this experiment, participants generated a novel category that did not occupy the same region of the feature space as the training categories, which means that the novel category cannot be viewed as a simple tweak of one of the existing categories.

Our discussion of Experiments 1 and 2 also pointed out that alternativeversions of the copy-and-tweak approach could be developed by departing from the underlying assumptions of exemplar models. Similarly, one could develop an alternative copy-and-tweak approach that would account for the data from Experiments 3 and 4. Consider, for example, a model that tweaks an entire category by sampling a distance and a dimension and simultaneously translating all category exemplars by the sampled distance along the sampled dimension. The resulting category would be superficially very different from the original, but would preserve the correlations between dimensions. The approach just described could account for our data without explicitly learning the correlations between dimensions, but additional assumptions would be needed to explain why translations over long distances qualify as acceptable tweaks.

8. General discussion

The sampling account of exemplar and category generation proposes that categories and exemplars are generated by sampling from probability distributions. We evaluated this account using two sets of behavioral experiments. The first two experiments explored exemplar generation by asking participants to generate exemplars of a structured category, and highlighted the fact that the sampling account naturally accommodates structured knowledge. The results were most consistent with a model that incorporated both graded probability distributions and structured representations. Experiments 3 and 4 explored category generation by asking participants to generate a new category that belonged to the same family as a set of example categories. The results were best accounted for by a model that represented knowledge both about the individual categories observed and the characteristics that unified them as a family.

Taken together, our results suggest that people find exemplar and category generation relatively natural, and that the sampling account can help to explain these abilities. Our data show that people are able to generate new exemplars and categories after observing a relatively small number of examples, a result that is consistent with previous psychological research on creative cognition. We now discuss the broader implications of this work and consider how it might be extended.

8.1. Mechanistic accounts of generation

A complete understanding of exemplar and category generation will require an account of the mechanisms involved. We described a mechanistic account that relies on a primitive cognitive mechanism analogous to rolling a weighted die. The sample-by-parts account proposes that people maintain separate distributions for the different parts of a category and generate novel exemplars by "rolling mental dice" in order to independently sample the parts of each exemplar. The sample-by-parts account is supported by our finding that participants in Experiments 1 and 2 tended to generate novel exemplars one part at a time.

The simplest version of sample-by-parts does not apply to Experiments 3 and 4 because the categories in those experiments included exemplars that varied along dimensions that were correlated rather than independent. We found, however, that participants tended to generate new categories either one dimension at a time or one crystal at a time. Both approaches are analogous to sample-by-parts, where the "parts" of a novel category are either the exemplars that belong to the category or the dimensions along which these exemplars vary.

Sample-by-parts is just one possible approach for developing mechanistic models of generation. Although this approach is unlikely to account for all of the contexts in which people generate new exemplars and categories, it serves to illustrate how the sampling account can be grounded in mechanistic models of generation. Future work can consider alternative mechanistic models and attempt to characterize a set of mechanisms that can collectively account for the full range of generation problems that humans are able to solve.

8.1.1. Beyond independent sampling

Another issue for future work on mechanistic models concerns the nature of the sampling itself. Both the sampling account and the sample-by-parts account assume that exemplars are sampled independently from a learned distribution. This simple assumption, however, is only a starting point, and other sampling strategies could be considered, such as choosing a set of exemplars that are representative in the sense that they provide good coverage of the category distribution (Kahneman & Tversky, 1972).

Our behavioral results from Experiment 2 suggest that at least some participants were not producing independent samples. Recall that in this experiment, exemplar (4,4) included the two rarest parts, leading the model to predict that this exemplar would be the least probable valid response. Participants, however, generated this exemplar about as often as several other valid category exemplars (see Fig. 7a), which is inconsistent with the assumption of independent sampling. An additional violation of independent sampling was observed in Experiments 3 and 4, in which participants generated sets of exemplars that spanned the full range of feature dimensions more often than would be predicted by independent sampling. Both of these results are consistent with goal-directed behavior, where the goal might be to generate a set of exemplars that provide the greatest possible coverage of the category, or to deliberately generate less probable exemplars with the intention of "balancing" feature frequencies, as several participants explicitly noted in their written explanations. Our experimental tasks were not intended to suggest such goals, but some participants may have adopted them anyway. Instead of sampling directly from exemplar and category distributions, these participants may have used alternative heuristics to generate sets of samples that were maximally representative of these distributions.

8.2. Generation and other inductive problems

The literature on categorization has focused overwhelmingly on classification (for a review, see Ashby & Maddox, 2005), but there are many other inductive problems that rely on category knowledge (Kemp & Jern, 2009). Markman and Ross (2003) have argued that classification is only one of many different ways in which people use knowledge about categories. For example, category knowledge is used for feature inference (Yamauchi & Markman, 2000; Yamauchi & Yu, 2008), problem solving (Ross, 1999), causal reasoning (Rehder, 2003; Rehder & Kim, 2006), and explanation (Lombrozo, 2009; Williams & Lombrozo, 2010). We propose that generation is another fundamental way in which people use their category knowledge.

There are important connections between some of these inductive problems. For example, studies of feature inference have generally considered settings in which a single feature is missing and must be inferred, but feature inference can also be studied in settings where multiple features are missing. From this perspective, exemplar generation can be considered a special case of feature inference in which a category label is provided and all features must be inferred (or generated). This observation suggests that a basic challenge for categorization research is to characterize how the full set of inductive problems are related to each other (Kemp & Jern, 2009).

Although problems like classification, feature inference, and generation appear to be related, it seems likely that some of these problems are inherently more challenging than others. For example, a magic square is an arrangement of integers from 1 to n in an $n \times n$ grid such that the sums of every row and every column are equal. While it is easy to verify whether a grid of numbers is a magic square (classifying an exemplar), it can be challenging to generate one, suggesting that generation may be the more difficult problem. There is some evidence for a proposal of this sort in the developmental literature. For example, infants are able to classify exemplars (Rakison & Yermolayeva, 2010), but the ability to imagine novel concepts does not appear to fully develop until late childhood (Berti & Freeman, 1997).

There are also theoretical reasons to believe that generation is more challenging than classification. For example, computer scientists have studied an extensive family of problems known as NP problems, and solutions to problems in this family are computationally "easy" to verify but are believed to be computationally "hard" to generate (Gasarch, 2002; Kleinberg & Tardos, 2005). In the psychological literature, Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) suggest that feature vectors require more bits to encode than category labels, and therefore argue that predicting feature vectors given category labels is intrinsically more difficult than predicting category labels given feature vectors.

8.3. Generative and discriminative models

Considering the relationships between inductive problems motivates the need for a unifying approach that can handle all of these problems. As discussed earlier, the sampling account belongs to a family of models known as generative models. In contrast, most existing psychological models of categorization are discriminative models.

The generative approach to categorization is appealing in part because it can handle multiple problems, including classification (our rating task results from Experiments 1 and 2; Anderson, 1991; Ashby & Alfonso-Reese, 1995; Fried & Holyoak, 1984), feature inference (Anderson & Fincham, 1996), and generation. A generative model provides a unified account of these problems because it learns a distribution $p(x,y)$ over exemplars x and category labels y that can be used for multiple purposes. This generality, however, may prevent the approach from accounting for important differences between inductive problems. Studies have demonstrated that what people learn about a category is shaped by the specific problem that they are facing (Love, 2005; Markman & Ross, 2003). For example, learners exposed to the same information can acquire different category representations depending on whether they are engaged in a classification task or a feature inference task (Markman & Ross, 2003; Sakamoto & Love, 2010), and this result is not captured by a generative model that uses the same distribution $p(x,y)$ for both tasks. In contrast, SUSTAIN is a discriminative model that accounts well for both classification and feature inference and successfully predicts that different category representations are acquired in these two tasks (Love et al., 2004).

The strengths and weaknesses of generative and discriminative models have been extensively discussed by psychologists (Hsu & Griffiths, 2010; Pothos & Bailey, 2009; Pothos et al., 2011; Wills & Pothos, 2012) and machine learning researchers (Bishop, 2006; Ng & Jordan, 2002; Xue & Titterington, 2008). Discussion on this topic continues, but we believe the evidence suggests that both approaches are valuable, and that human learning can be generative or discriminative, depending on the context (Hsu & Griffiths, 2009, 2010). For example, feature inference problems (Anderson & Fincham, 1996; Yamauchi & Markman, 1998) are naturally handled by generative models, because generative models can use the learned exemplar distribution $p(x|y)$ to fill in missing features of observed objects. On the other hand, problems involving ideal categories (Barsalou, 1985) are naturally handled by discriminative models, because the best examples of an ideal category do not fall near the peak of the exemplar distribution. For example, most diet foods have calorie counts greater than zero, but the best examples of diet foods have zero calories.

As we have shown, one limitation of discriminative models is that they do not account for exemplar and category generation. Existing discriminative models can be modified to account for generation, but the two such modifications that we considered (randomly-sample-and-score and copy-and-tweak) did not account for our data as well as the generative sampling-based account. These modified discriminative models, however, may account better for generation in other contexts. For example, our experiments used a paradigm where the stimuli were visible at all times, but if the stimuli had been presented sequentially, participants might have been inclined to generate novel exemplars by copying and tweaking the last exemplar that they observed.

Our claim, then, is not that generative models provide a complete account of categorization, or even that these models provide a complete account of exemplar and category generation. Instead, we suggest that probabilistic generative models are useful for understanding how category knowledge is learned and used in some but not all contexts. At present there is no theoretical approach that accounts for classification, feature inference, and generation, and that explains how task context shapes learning in all of these settings. Developing an approach that meets all of these criteria is an important challenge for future work.

8.4. Generation and task context

Although a fully general account of categorization may take years to develop, exploring how task context influences generation is an immediate priority for future work. Our generation tasks consisted of two phases, a category learning phase and a generation phase, and task context might affect either phase. Manipulations that affect the category learning phase are familiar from previous work. For example, in all of our experiments, the training stimuli were visible at all times in order to minimize memory demands, but presenting these stimuli one by one may change what participants learn about the underlying categories. If the training stimuli are presented in sequence, past studies suggest that learners who make a feature inference about each stimulus are more likely to acquire accurate category distributions than learners who make a classification judgment about each stimulus (Markman & Ross, 2003; Sakamoto & Love, 2010). By shaping the category knowledge that learners acquire, experimental manipulations of this kind will affect performance on any subsequent generation task.

After learning about a category, participants could be asked to generate novel exemplars in many different ways. For example, participants could be asked to generate the most typical novel exemplar that they can imagine, to generate two novel exemplars that are as different as possible, or to generate an exemplar that is as unusual as possible while still qualifying as a valid instance of the category. The probability distribution learned by the sampling account can be used to address all of these tasks. For example, an atypical exemplar might be generated by finding an exemplar that is as far as possible from the category mean but still exceeds some threshold probability. People might also be asked to generate exemplars that satisfy certain constraints. For example, Ward (1994) asked participants to imagine an animal that lived on a planet whose surface consisted mainly of molten rock. The sampling account may be able to handle tasks of this kind by sampling from conditional probability distributions. A conditional distribution of this kind could be created by beginning with the standard exemplar distribution over animals and then conditioning on the fact that the novel animal must be able to survive on a molten planet. Future work, however, is needed to turn these speculations into concrete model predictions and to evaluate them against behavioral data.

9. Conclusion

The generation of novel concepts, objects, and ideas occurs in settings from the mundane (making a new dish for dinner) to the exceptional (conceiving of a previously unobserved species). This paper argued that computational models of categorization should aim to account for generation in addition to more commonly studied problems such as classification and feature inference. Most psychological models of categorization do not account for generation, but we developed and evaluated a sampling account that maintains probability distributions over exemplars and categories and generates novel items by sampling from these distributions. Our experiments explored generation in the context of categorization, but the proposal that people sample from probability distributions seems relatively general and may help to account for creative behavior in other settings. Characterizing the computational basis of creativity is obviously a challenging problem, but generative probabilistic models may provide part of the solution.

Acknowledgements

Preliminary data from Experiments 1 and 2 were presented at the 31st Annual Conference of the Cognitive Science Society. We thank Faye Han for helping to collect the data for Experiments 1 and 2 and Meghan Reese and Jessica Lee for helping to collect the data for Experiments 3 and 4. We thank Art Markman and three anonymous reviewers for feedback on the manuscript, and John Anderson and David Rakison for feedback throughout the development of this work. This work was supported by NSF Grant CDI-0835797 and the Pittsburgh Life Sciences Greenhouse Opportunity Fund.

Appendix A. Modeling details

A.1. The sampling account: Averaging over category parameters

In the main text, we described the sampling account of exemplar and category generation as involving discrete steps in which the category (or higher-order category) parameters are first inferred from the training data and subsequently used to generate new exemplars or categories. These steps, however, can be combined in order to directly estimate the full exemplar and category distributions.

A.1.1. Exemplar generation

The exemplar generation distribution, $p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y})$, in Eq. (2) can be computed directly by integrating out the category parameters $\boldsymbol{\theta}_{\tilde{y}}$:

$$p(\tilde{x}|\tilde{y},\mathbf{x},\mathbf{y}) = \int_{\boldsymbol{\theta}_{\tilde{y}}} p(\tilde{x}|\boldsymbol{\theta}_{\tilde{y}})p(\boldsymbol{\theta}_{\tilde{y}}|\mathbf{x},\mathbf{y})d\boldsymbol{\theta}_{\tilde{y}}.$$

The resulting distribution is referred to as the posterior predictive distribution in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 2004). The second term in the integral can be computed by applying Bayes rule:

$$p(\boldsymbol{\theta}_{\tilde{y}}|\mathbf{x},\mathbf{y}) \propto p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}_{\tilde{y}})p(\boldsymbol{\theta}_{\tilde{y}}).$$

If we assume that exemplars of a category are independent samples, this expression can be rewritten as

$$p(\boldsymbol{\theta}_{\tilde{y}}|\mathbf{x},\mathbf{y}) \propto \left(\prod_{\{i:y_i=\tilde{y}\}} p(x_i|\boldsymbol{\theta}_{\tilde{y}})\right)p(\boldsymbol{\theta}_{\tilde{y}}). \tag{A.1}$$

In the product above, we assume that only exemplars that belong to category $\tilde{y}$ depend on that category's parameters. Eq. (A.1) indicates that modeling exemplar generation requires specifying two distributions: a prior probability distribution over the category parameters, $p(\boldsymbol{\theta}_{\tilde{y}})$, and a likelihood distribution, $p(x_i|\boldsymbol{\theta}_{\tilde{y}})$, that specifies how exemplars are generated.

A.1.2. Category generation

The category generation distribution $p(\boldsymbol{\theta}_{\tilde{y}}|\mathbf{x},\mathbf{y})$ can be also be computed by integrating out both the training category parameters $\boldsymbol{\theta}$ and the higher-order category parameters $\boldsymbol{\phi}$:

$$p(\boldsymbol{\theta}_{\tilde{y}}|\mathbf{x},\mathbf{y}) = \int_{\phi}\int_{\theta} p(\boldsymbol{\theta}_{\tilde{y}}|\boldsymbol{\phi})p(\boldsymbol{\phi},\boldsymbol{\theta}|\mathbf{x},\mathbf{y})d\boldsymbol{\theta}d\boldsymbol{\phi},$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\boldsymbol{\theta}_j$ denotes the parameters for category *j*. The second term in the integral can be computed by applying Bayes' rule:

$$\begin{aligned} p(\boldsymbol{\phi},\boldsymbol{\theta}|\mathbf{x},\mathbf{y}) &\propto p(\mathbf{x},\mathbf{y}|\boldsymbol{\phi},\boldsymbol{\theta})p(\boldsymbol{\phi},\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}) \\ &= \left(\prod_{j=1}^{m}\left(\prod_{\{i:y_i=j\}} p(x_i|\boldsymbol{\theta}_j)\right)p(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right)p(\boldsymbol{\phi}). \end{aligned} \tag{A.2}$$

We assume here that category parameters $\boldsymbol{\theta}_j$ are conditionally independent given the higher-order category parameters $\boldsymbol{\phi}$.

Eq. (A.2) indicates that modeling category generation requires specifying three distributions. The first is a prior probability distribution over the higher-order category parameters $p(\boldsymbol{\phi})$, and the remaining two are likelihood distributions: $p(\boldsymbol{\theta}_j|\boldsymbol{\phi})$ specifies how category parameters are generated and $p(x_i|\boldsymbol{\theta}_j)$ specifies how exemplars are generated.

A.2. Models for Experiments 1 and 2

This section gives a complete specification of the exemplar generation models tested in Experiments 1 and 2.

A.2.1. Structured sampling model

The structured sampling model generates exemplars by sampling from the exemplar generation distribution, $p(\tilde{x}|S,\eta_1,\ldots,\eta_{|S|})$, where S denotes the category structure and η_i specifies the distribution over parts for slot *i*. We assume that the structure S is sampled from a uniform distribution over the 15 possible structures, that each distribution η_i is drawn from a Dirichlet prior with parameter α, and that part x_i is sampled from multinomial distribution η_i:

$$\begin{aligned} S &\sim \text{Uniform}(\{s_1,\ldots,s_{15}\}) \\ \eta_i|S &\sim \text{Dirichlet}(\alpha) \\ x_i|\eta_i &\sim \text{Multinomial}(\eta_i). \end{aligned}$$

We assume that the distribution η_i is defined over all possible permutations of letters that could fill slot *i*. For example, if the slot has *m* feature positions and there are *k* possible letters, then there are k^m possible parts that could fill the slot. We set the parameter α by assuming that parts are expected to recur, and that the probability of observing a part in a given slot of a category exemplar after previously observing that part in the same slot of another category exemplar is 0.5. Anderson's (1991) rational model of categorization makes a related assumption, and refers to the probability of 0.5 as a "coupling probability." Setting the coupling probability to 0.5 implies 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.

For the rating task, we assumed that all of the rating items were sampled uniformly from the full feature space. This means that when computing the classification distribution,

$$p(\tilde{y}|\tilde{x}) = \frac{p(\tilde{x}|\tilde{y})p(\tilde{y})}{p(\tilde{x})},$$

the term $\frac{p(\tilde{y})}{p(\tilde{x})}$ is a constant and the ratings are proportional to the likelihoods of the exemplar generation distribution, $p(\tilde{x}|\tilde{y})$.

A.2.2. Rule-based model

The rule-based model is equivalent to the structured sampling model if each slot distribution is assumed to be a uniform distribution over the parts that have been previously observed in that slot. In other words, let $\mathbf{x}_{i|S}$ be the set of observed parts that fill slot i given structure S in the training data. The rule-based model assumes that each part x_i is sampled from a uniform distribution over this set:

$$\begin{aligned} S &\sim \text{Uniform}(\{s_1, \ldots, s_{15}\}), \\ x_i|S &\sim \text{Uniform}(\mathbf{x}_{i|S}). \end{aligned}$$

A.2.3. Independent features model

The independent features model is equivalent to the structured sampling model if the structure S is assumed to consist of four independent features, as shown in the far right of Fig. 4a. Let this structure be denoted by s_{15}. Then the generative assumptions of the independent features model can be expressed as follows:

$$\begin{aligned} S &= s_{15}, \\ \eta_i|S &\sim \text{Dirichlet}(\alpha), \\ \tilde{x}_i|\eta_i &\sim \text{Multinomial}(\eta_i). \end{aligned}$$

The parameter α was set in the same way as for the structured sampling model.

The predictions for all three sampling-based models were produced by drawing 100,000 samples from each model's exemplar generation distribution.

A.2.4. The RSS model

The RSS model samples an exemplar from a uniform distribution over possible exemplars, scores it using the classification distribution, and then generates the exemplar with a probability that depends on its score. It repeats this procedure until three exemplars have been generated.

The model can be formally specified as follows. First an exemplar $\tilde{x}_s$ is sampled from the space of possible exemplars:

$$\tilde{x}_s \sim \text{Uniform}(\{1 : k^4\}),$$

where k is the number of observed letters and the exponent comes from the fact that there are four feature positions. Let $c(\tilde{x}_s) = p(\tilde{x}_s|\tilde{y}, \mathbf{x}, \mathbf{y})$ be the score assigned to the exemplar, based on the classification distribution. The probability of generating exemplar $\tilde{x}_s$ is

$$P(\text{generate } \tilde{x}_s) = (1 - \xi) \cdot \gamma \cdot c(\tilde{x}_s) + \xi,$$

where $\gamma = \frac{1}{\max_{\tilde{x}} c(\tilde{x})}$ (the maximum is taken over all possible exemplars so that the exemplars with the highest score are generated with probability 1) and ξ is the baseline generation probability. ξ was chosen so that the expected number of samples needed to generate three exemplars was approximately equal to a parameter $E(N)$. Fig. 9 shows model predictions for two values of $E(N)$ for Experiments 1 and 2. The values in the figure are approximate. The mean numbers of samples needed in our simulations to produce the predictions in Fig. 9a were 949 and 16,492. The mean numbers of samples needed to produce the predictions in Fig. 9b were 968 and 41,531. The model predictions were produced by using the model to generate 10,000 sets of three exemplars.

A.2.5. The batch RSS model

An alternative version of the RSS model might sample and score multiple exemplars before making a decision about which exemplars to generate. One way to make this decision is to generate the three exemplars with the highest scores. This approach might make reasonable predictions for Experiment 1 but cannot account for Experiment 2. A model using this approach would predict that the three most probable exemplars (exemplars (1,1), (2,2), and (3,3) in Fig. 5) should be disproportionately favored over the other valid exemplars. If the sample size is sufficiently large, then such a model would generate only these three exemplars.

We therefore evaluated a model that generates samples probabilistically in proportion to their scores. We again introduced a sample size parameter, N. This time, the sample size parameter specifies the number of samples that are drawn and scored. First, N exemplars are sampled without replacement from the space of possible exemplars. Let the ith exemplar be denoted $\tilde{x}_{s_i}$. As for the sequential RSS model, let $c(\tilde{x}_{s_i}) = p(\tilde{x}_{s_i}|\tilde{y}, \mathbf{x}, \mathbf{y})$ be the score assigned to exemplar $\tilde{x}_{s_i}$, based on the classification distribution. Exemplars are generated according to a Multinomial distribution based on the scores:

$$\tilde{x} \sim \text{Multinomial}([R \cdot c(\tilde{x}_{s_1}), \ldots, R \cdot c(\tilde{x}_{s_N})]),$$

where $R = 1/\Sigma_i c(\tilde{x}_{s_i})$ is a normalization constant.

Model predictions for Experiments 1 and 2 are shown in Fig. A.18. The predictions were produced by using the model to generate 2,000,000 sets of three exemplars. The figure shows predictions for two settings of the sample size parameter. As for the RSS model in the main text, the batch RSS model is provably equivalent to the structured sampling model when the parameter N is set sufficiently high. Fig. A.18 shows that the model does not account for our data when $N = 1000$. Reducing the sample size further would only make the fit of the model worse, and we therefore conclude that the batch RSS model does not account for our data when the sample size parameter is set to a psychologically plausible value.

A.2.6. The copy-and-tweak model

The copy-and-tweak model uses similarity to observed exemplars as the basis for both classification and generation. We used a similarity rule derived from the GCM (Nosofsky, 1986) that defines the similarity s between exemplars x_1 and x_2 as $s(x_1,x_2) = \exp(-c \cdot d(x_1,x_2))$, where c is a scaling parameter and $d(x_1,x_2)$ is the distance between the two exemplars. We set c to 1 and used the Hamming distance, which counts the number of features for which the two exemplars have different values.

To derive classification predictions from the copy-and-tweak model for the rating task, we summed the similarities of each rating exemplar x_r to each of the training exemplars x_i: $\text{rating}(x_r) = \sum_i s(x_r, x_i)$. We then applied a z-score transformation to produce the predictions in Fig. 10.

To derive generation predictions from the copy-and-tweak model, we computed similarity ratings for all possible exemplars and drew 2,000,000 samples from a Multinomial distribution based on the ratings:

$$\tilde{x} \sim \text{Multinomial}([R \cdot \text{rating}(\tilde{x}_1), \ldots, R \cdot \text{rating}(\tilde{x}_n)]),$$

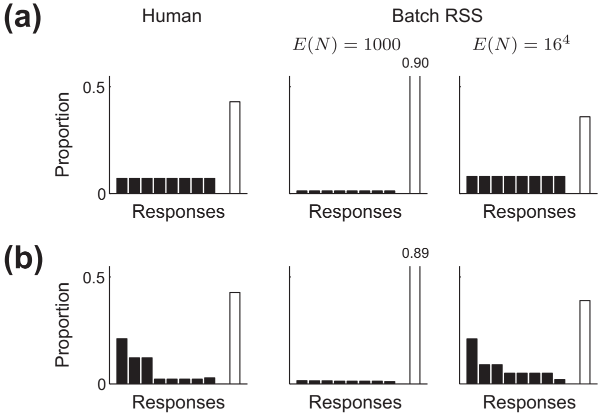

Fig. A.18. The batch RSS model makes different predictions depending on the value of the sample size N. The third column shows the model's predictions when the entire space of 16^4 exemplars was scored.

where $R = 1/\Sigma_i \text{rating}(\tilde{x}_i)$ is a normalization constant. The reported predictions are based on a model which considers all possible exemplars that can be constructed from the 16 observed letters ($n = 16^4$). We also implemented an alternative model that considers only exemplars in which each letter can only appear in positions where it has been previously observed ($n = 256$). The two models produced similar predictions, but the former version performed slightly better.

A.3. Experiment 3 and 4 models

This section gives complete specifications of the category generation models tested in Experiments 3 and 4, and describes how the predictions of these models were computed.

A.3.1. The hierarchical sampling model

Recall that we represent each category y by a three-dimensional normal distribution with a mean μ_y and a covariance matrix Σ_y. That is, the category parameters $\boldsymbol{\theta}_y = (\mu_y, \Sigma_y)$. The higher-order category parameters specify a "template" distribution from which categories are generated, $\boldsymbol{\phi} = (\mu_0, \Sigma_0)$. See Fig. 12 for a graphical illustration of these assumptions.

The hierarchical sampling model assumes that categories are generated by starting with the higher-order category template and randomly tweaking the values of its mean μ_0 and the cells of its covariance matrix Σ_0. To be more precise, let Σ_0 be the following matrix:

$$\Sigma_0 = \begin{bmatrix} \sigma_{0_{11}} & \sigma_{0_{12}} & \sigma_{0_{13}} \\ \cdot & \sigma_{0_{22}} & \sigma_{0_{23}} \\ \cdot & \cdot & \sigma_{0_{33}} \end{bmatrix}.$$

We have omitted the cells below the main diagonal because covariance matrices must be symmetric (i.e., $\sigma_{0_{ij}} = \sigma_{0_{ji}}$). The category parameters can be inferred by using Eq. (A.2) and making the following generative assumptions:

$$\begin{aligned} p(\mu_0) &\propto 1, \\ p(\sigma_{0_{ij}}) &\propto 1, \\ \mu_y &\sim \text{Normal}(\mu_0, \sigma_\mu), \\ \sigma_{y_{ji}} = \sigma_{y_{ij}} &\sim \text{Normal}(\sigma_{0_{ij}}, \sigma_\Sigma). \end{aligned}$$

Here σ_μ and σ_Σ are parameters that represent how much cross-category variability is expected for category means and covariance matrices, respectively. For all results in this paper, we set $\sigma_\mu = 0.5$ and $\sigma_\Sigma = 0.05$ to capture the idea that category means are expected to vary much more than category covariances. The above assumptions are also subject to the implicit constraint that the resulting covariance matrix Σ_y must be positive-definite. In our simulations, this constraint was applied by discarding sets of samples that did not produce a positive-definite matrix.

Because the feature values of the category exemplars were restricted to the [0,1] interval, we applied an additional assumption that the generated exemplars are mapped to this range by a logistic transformation. We therefore assume that each exemplar x from category y is generated as follows:

$$\begin{aligned} x' &\sim \text{Normal}(\mu_y, \Sigma_y), \\ x &= \frac{1}{1 + e^{-x'}}. \end{aligned}$$

The model assumes that this transformation has already been applied to the observed training data, and therefore applies an inverse transformation to the data before learning the category generation distribution.

A.3.2. The independent categories model

The independent categories model does not learn any category constraints and instead assumes that all categories are generated independently. Formally, the model assumes that all category parameters are sampled from uniform distributions, again subject to the constraint that the resulting covariance matrices must be positive-definite:

$$\begin{aligned} p(\mu_y) &\propto 1, \\ p(\sigma_{y_{ij}}) &\propto 1, \\ x' &\sim \text{Normal}(\mu_y, \Sigma_y), \\ x &= \frac{1}{1+e^{-x'}}. \end{aligned}$$

A.3.3. Model predictions

We used a Metropolis–Hastings sampler with Gaussian proposal distributions to obtain estimates of the category generation distributions with the hierarchical sampling model. To simplify the inference, we used maximum likelihood estimates of the training category means and covariance matrices instead of estimating distributions over these parameters. In other words, we produced samples from a probability distribution only over the higher-order category parameters μ_0 and Σ_0. We generated 150,000 samples from the posterior distribution on μ_0 and Σ_0 after a burn-in period of 50,000 samples. To implement the independent categories model, we generated 50,000 samples from the prior distribution on μ_0 and Σ_0.

We generated 50,000 categories of six exemplars for each model. Each category was generated by first sampling a pair of higher-order category parameters from the category generation distribution. The category parameters and individual exemplars were subsequently generated by applying the generative assumptions shown in the model specifications above.

References

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

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 259–277.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481.

Ashby, F. G., & Gott, R. F. (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., & 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. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660.

Barsalou, L. W., & Prinz, J. J. (1997). Mundane creativity in perceptual symbol systems. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 267–307). Washington, DC: American Psychological Association.

Berti, A. E., & Freeman, N. H. (1997). Representational change in resources for pictorial innovation: A three-component analysis. *Cognitive Development, 12*, 501–522.

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

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*, 347–356.

Cobb, S. W., & Willems, M. (1996). *The Brown Derby Restaurant: A Hollywood Legend*. Rizzoli International Publications.

Colunga, E., & Smith, L. B. (2008). Flexibility and variability: Essential to human cognition and the study of human cognition. *New Ideas in Psychology, 26*, 174–192.

Daeschler, E. B., Shubin, N. H., & Jenkins, F. A. Jr., (2006). A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature, 440*, 757–763.

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

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

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14*, 119–130.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–257.

Gasarch, W. I. (2002). The P=?NP poll. *SIGACT News, 33*, 34–47.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Florida: CRC Press.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.

Hoyer, P .O., & Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems* 16.

Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems* 22.

Hsu, A., & Griffiths, T. L. (2010). Effects of generative and discriminative learning on use of category variability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychnomic Bulletin & Review, 15*, 256–271.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62*, 499–516.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition, 34*, 57–83.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65*, 103–135.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Kemp, C., & Jern, A. (2009). A taxonomy of inductive problems. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307–321.

Kleinberg, J., & Tardos, E. (2005). *Algorithm design*. Addison Wesley.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 92*, 22–44.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. New York: Cambridge University Press.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition, 110*, 248–253.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14*, 195–199.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.

Luce, R. D. (1959). *Individual choice behavior*. John Wiley.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development, 6*, 17–46.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*, 592–613.

Marsh, R. L., Landau, J. D., & Hicks, J. L. (1996). How examples may (and may not) constrain creativity. *Memory & Cognition, 24*, 669–680.

Marsh, R. L., Ward, T. B., & Landau, J. D. (1999). The inadvertent use of prior knowledge in a generative cognitive task. *Memory & Cognition, 27*, 94–105.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences, 108*, 12491–12496.

Ng, A., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of Neural Information Processing Systems* 14.

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

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychnomic Bulletin & Review, 5*, 345–369.

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

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Perfors, A., & Tenenbaum, J. B. (2009). Learning to learn categories. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1062–1080.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303–343.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., et al (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition, 121*, 83–100.

Pothos, E. M., & Wills, A. J. (Eds.). (2011). *Formal approaches in categorization*. Cambridge, MA: Cambridge University Press.

Rakison, D. H., & Yermolayeva, Y. (2010). Infant categorization. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 894–905.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*, 909–957.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science, 27*, 709–748.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 659–683.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10*, 759–784.

Rehling, J. A. (2001). Letter spirit (part two): Modeling creativity in a visual domain. Ph.D. thesis Indiana University.

Rehling, J. A., & Hofstadter, D. R. (2004). Letter spirit: A model of visual creativity. In M. Lovett, C. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the sixth international conference on cognitive modeling* (pp. 249–254). Mahwah, NJ: Lawrence Erlbaum.

Ross, B. H. (1999). Postclassification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 743–757.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology, 46*, 178–210.

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied, 16*, 361–377.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167.

Shelton, J. A., Bornschein, J., Sheikh, A.-S., Berkes, P., & Lücke, J. (2011). Select and sample—A model of efficient neural inference and learning. In *Advances in Neural Information Processing Systems* 24.

Shi, L., & Griffiths, T. L. (2009). Neural implementation of Bayesian inference by importance sampling. In *Proceedings of Neural Information Processing Systems* 22.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*, 13–19.

Smith, S. S., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition, 21*, 837–845.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1–26.

Tenenbaum, J.B. (1999). Bayesian modeling of human concept learning. In *Proceedings of Neural Information Processing Systems* 11.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273–13278.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53*, 27–58.

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

Ward, T. B. (1995). What's old about new ideas? In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach*. Cambridge, MA: MIT Press.

Ward, T. B., Patterson, M. J., & Sifonis, C. M. (2004). The role of specificity and abstraction in creative idea generation. *Creativity Research Journal, 16*, 1–9.

Ward, T. B., Patterson, M. J., Sifonis, C. M., Dodds, R. A., & Saunders, K. N. (2002). The role of graded category structure in imaginative thought. *Memory & Cognition, 30*, 199–216.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science, 34*, 776–806.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin, 138*, 102–125.

Xue, J.-H., & Titterington, D. M. (2008). Comment on "On discriminative vs.generative classifiers: A comparison of logistic regression and naive Bayes". *Neural Processing Letters, 28*, 169–187.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 776–795.

Yamauchi, T., & Yu, N.-Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition, 36*, 544–553.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*, 301–308.