Creating Something Different:
Similarity, Contrast, and Representativeness in Categorization

Joseph L. Austerweil[1+] Shi Xian Liew[1+], Nolan Conaway[1], and Kenneth J. Kurtz[2]

[1]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Psychology, Binghamton University, Binghamton, NY, USA

Author Note
Correspondence concerning this article should be addressed to: Shi Xian Liew, 1202 West Johnson Street, Madison, WI 53706. E-mail: shixianliew@gmail.com
[+]These authors share equal contribution to this work.

Abstract

The ability to generate new concepts and ideas is among the most fascinating aspects of human cognition, but we do not have a strong understanding of the cognitive processes and representations underlying concept generation. In this paper, we study the generation of new categories using the computational and behavioral toolkit of traditional artificial category learning. Previous work in this domain has focused on how the statistical structure of known categories generalizes to generated categories, overlooking whether (and if so, how) contrast between the known and generated categories is a factor. We report three experiments demonstrating that contrast between what is known and what is created is of fundamental importance for categorization. We propose two novel approaches to modeling category contrast: one focused on exemplar dissimilarity and another on the representativeness heuristic. Our experiments and computational analyses demonstrate that both models capture different aspects of contrast's role in categorization.

Keywords: categorization; concepts; category learning; generation; computational modeling; exemplar models; representativeness

1 Introduction

Creating new ideas is one of the most fascinating and important human capabilities. For example, cell phones and smart phones are newly created categories of objects that young and old adults learned and have become reliant on in the last few decades. Computational researchers have identified generating novel objects and objectives as one of the defining, and most difficult to formalize, characteristics of intelligent life (Lake, Ullman, Tenenbaum, & Gershman, 2017; Lehman & Stanley, 2011; Taylor et al., 2016). Yet, the cognitive mechanisms that enable people to innovate are not well understood and are understudied by scientists. In part, this is due to the inherent difficulty of designing and conducting experiments that test these capabilities. How would a scientist devise an experiment that induces a participant to create new categories that are as interesting as cell phones and smart phones? Devising such an experiment is still beyond our capability. Instead, we take a step towards this goal, and investigate the fundamental constraints and expectations that people have when they generate new categories.

Creating new concepts are, however, not altogether different from the types of behaviors typically studied in cognitive psychology laboratories. In particular, generating a member of a novel class can be considered a ‘special case’ application of existing category knowledge (Kemp & Jern, 2014; Kurtz, 2015). Research in categorization typically focuses on what properties of a category affect human learning of an object’s category given its features (Kurtz, Levering, Stanton, Romero, & Morris, 2013; Shepard, Hovland, & Jenkins, 1961), or the prediction of an object’s unobserved features given some of its other features and/or its category (Markman & Ross, 2003). The generation of members of a new category consists of inferring *all* features for a *novel* category label. Thus, we can make progress formalizing the processes involved in category generation by extending theories of categorization to this case. This will also serve to test the breadth of the explanatory power of categorization models to tasks beyond typical category learning.

Previous work in category learning and generation has established that people are highly sensitive to the structural properties of categories, such as correlations between the features of category members and the relation between items within the same category and those in different categories (Regier, Kay, & Khetarpal, 2007; Rosch & Mervis, 1975; Shepard et al., 1961; S. M. Smith, Ward, & Finke, 1995).[1] Inspired by this work, previous research on the topic of category generation has explored a similar principle: People tend to create new categories that have similar *statistical regularities* as previously learned categories (Jern & Kemp, 2013; Ward, 1994). These statistical regularities reflect general structural patterns across old and new categories and are not restricted to strictly quantitatively similar properties.

Although statistical regularity is an important characteristic of generating new categories, it cannot be the only one. Taken to the extreme, the best "new" category in terms of having the same statistical regularities to other categories would be identical to a known category that is representative of the domain (and thus, not new at all). Contrast from other known categories should play a role. Indeed, scientists in other fields, such as marketing (Berger, 2016) and sociology (Rogers, 2003), highlight the critical importance of contrast in the creation of new ideas.

To successfully generate something novel, what is generated must be different from what is already known. This fundamental constraint, "being different", or contrasting from other categories in the relevant domain, is the focus of our work. Intuitively, the influence of contrast may be expected in a number of real-life scenarios involving the generation of new categories. For instance, a college student who has been eating instant noodles for a week may be more inclined to seek out something healthier and less carbohydrate-heavy (e.g., a Caesar salad) compared to something similar (a different flavor of instant noodles). An actor who is dissatisfied with being typecast as a villain may seek out roles as heroes as opposed to playing other antagonistic characters. An especially compelling example can be

[1] It is worth noting that sensitivity to feature correlations is not universal and can be dependent on factors such as the salience of correlations and task demands (e.g., see Chin-Parker & Ross, 2002; Malt & Smith, 1984; G. Murphy, 2004).

seen in Askin and Mauskapf (2017) who found that new songs that were more different to their peers were generally more popular than new songs that were similar to others. It appears that once there is the motivation to generate something new, there is an associated motivation to generate and observe something *different*.

Although implicitly assumed in some work, this constraint has been overlooked in previous research: To our knowledge, there has not been any systematic investigation addressing how generated categories *differ* from what is already known. Although the idea of category contrast is discussed throughout the categorization literature, and extends to a variety of other fields (e.g., color; Regier et al., 2007), the idea that a new category should be "different" is vague, as there are many ways it could be different from a previously observed category.

We examine two different definitions of how categories should differ from one another: exemplar dissimilarity and representativeness of the alternative category. To formalize the first technique, we build on the largely successful exemplar modeling framework (Medin & Schaffer, 1978; Nosofsky, 1984, 1986). In doing so, we propose a novel exemplar model of category generation, *Producing Alike and Contrasting Knowledge using Exemplar Representations* (PACKER), formalizing how new categories should differ from previous categories. It embodies the exemplar dissimilarity principle by incorporating a factor that repulses members of opposite categories from one another.

The second hypothesis for how contrast might affect categorization is contrast as *representativeness*. The representativeness heuristic (Kahneman & Tversky, 1972) states that people make judgments regarding an outcome based on how representative it is of the evidence given in the current context. The heuristic is a powerful theoretical construct that has been used to capture a range of complex patterns of human judgments (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974, 1983), especially those that deviate from normative theory (as given by a straightforward application of Bayes' rule). For example, the coinflip sequence TTHTH is perceived to be more random (due to it being representative of a coinflip sequence) than TTTTT (despite both having the same probability of having been generated from a fair coin). Although there is a healthy debate as to whether perceived randomness *really* deviates from normative theory (Griffiths, Daniels, Austerweil, & Tenenbaum, 2018; Hahn & Warren, 2009), the representativeness heuristic remains one of the main explanations of human judgments (Reimers, Donkin, & Le Pelley, 2018). To formalize the hypothesis that people expect a new category to be representative of the opposite of the current category, or contrast as *representativeness*, we use the Bayesian formulation of representativeness (Tenenbaum & Griffiths, 2001), which has been used successfully to capture human performance in color categorization (Abbott, Griffiths, & Reiger, 2016), image category learning (Abbott, Heller, Ghahramani, & Griffiths, 2011), and language grammar learning (Rafferty & Griffiths, 2010). We do so by extending the state-of-the-art category generation model (Jern & Kemp, 2013) and find that human category generation can also be captured as generating representative, rather than probable, samples.

The outline of the article is as follows. First we describe previous computational formalizations of theories of category generation and empirical work investigating them. We then present two hypotheses for how contrast might affect categorization and formalize them in computational models. The first is a novel exemplar model, which is designed to generate categories that systematically differ from existing categories in the domain. The second assumes that the goal of category generation is to create representative samples of the opposite of the observed categories. We present two experiments demonstrating strong and systematic effects of category contrast on concept generation, and we qualitatively and quantitatively analyze the behavioral results and the performance of each model in capturing human category generation. A third experiment is performed to highlight the differences between the two contrast models. We conclude with a discussion of the implications of our results for categorization and directions for future work.

2 Prior work

Much of what we know about concept generation and contrast comes from the foundational literature on creative cognition. In a series of reports, Ward and colleagues (Marsh, Ward, & Landau, 1999; S. M. Smith, Ward, & Schumacher, 1993; Ward, 1994, 1995; Ward, Patterson, Sifonis, Dodds, & Saunders, 2002) established that category generation is highly constrained by prior knowledge: Generated categories tend to consist of features observed in known categories, and they tend to exhibit the distributional properties found in known categories. In a seminal study, Ward (1994) asked participants to generate new species of alien animals by drawing and describing members of the species. People tended to generate species with the same features as on Earth (e.g., eyes, legs, wings), and possessing the same feature correlations as on Earth (e.g., feathers co-occur with wings). Likewise, aliens drawn from the same species tended to share more features with one another compared to members of opposite species.

The broader set of observations made by Ward and colleagues provide a great deal of insight into the role of prior knowledge in constraining category generation. Much of the work from this area (e.g., Marsh et al., 1999; S. M. Smith et al., 1993) focuses on how information provided to participants (such as an example of a species generated by other participants) can drastically diminish the difference of a new category from pre-existing categories. Theoretical accounts of these effects have primarily been grounded within the categorization literature. For example, the predominant "Path of Least Resistance" account (see Ward, 1994, 1995; Ward et al., 2002) proposes that, when generating a new species of animal, people retrieve from memory a known subcategory of animals (e.g., *bird*, *dog*, *horse*), and simply change some of the features to make something new. People are thought to change only features that are not characteristic of the retrieved category (e.g., if *bird* was retrieved, the presence of *wings* would not change, but *color* might). This theory incorporates elements of the highly influential basic-level categories framework (Rosch, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), as well as the exemplar view

(Brooks, 1978; Medin & Schaffer, 1978). While this work has been incredibly useful in providing a conceptual sketch of generation theories, its qualitative nature and the hand-drawn responses used in the experiments paradigms precludes the development of formal approaches that can be used to test them in a fine-grained manner.

Jern and Kemp (2013) recently showed that concept generation could be studied in a more controlled manner through the well-developed methods of an artificial categorization paradigm (see Kurtz, 2015, for a review). In Experiments 3 and 4 of their article, participants were exposed to members of experimenter-defined categories of ”crystals” varying in size, hue, and saturation. Following a training phase during which the experimenter-defined categories were learned, participants were asked to generate novel categories of crystals. In a finding mirroring Ward (1994), Jern and Kemp (2013) found that participants generated categories with the same distributional properties as the experimenter-defined categories. For example, after learning categories with a positive correlation between the size and saturation features (larger sized crystals were more saturated), participants generated novel categories with the same positive correlation. By replicating Ward (1994), they demonstrated that category generation can be studied in a well-known and highly controlled experimental paradigm.

The authors evaluated the predictions of several formal models on their data. Most notably, they showed that a hierarchical Bayesian model provided the strongest account of their results. Their model views observed examples as samples from an underlying category distribution, describing the location of the category in the space, as well as how it varies along each feature. In turn, each category is viewed as a sample from an underlying *domain* distribution, specifying distributional commonalities among the observed categories. Generated categories are thought to stem from the same domain distribution as observed categories, thus the distributional properties of observed categories will be preserved within the generated category.

Jern and Kemp (2013) additionally tested a “copy-and-tweak” model that broadly resembles the earlier “Path of Least Resistance” account. The core proposal is that participants generate new items by copying stored examples from memory and tweaking them to generate something new. The copy-and-tweak model differs from the Path of Least Resistance account in that it notably omits the hierarchical organization of categories, as well as selectivity to which features are changed (both of which are factors in the Path of Least Resistance account; Ward et al., 2002). Instead, their copy-and-tweak model corresponds to a direct exemplar-similarity approach (e.g., Nosofsky, 1984, 1986), generating new items according to their similarity to known members of the target category. The copy-and-tweak model provided a poor account of their results, as the experiments devised by Jern and Kemp were specifically designed to challenge it.

It is worth pointing out that the early observation of regularities in distributional properties across categories is not confined to the work of Ward and colleagues. Most notably, Thomas (1998) trained participants to classify circles of different sizes, each with a radial line of varying orientation, into two categories. During a surprise prediction phase, participants were asked to generate a value for a missing feature for a certain category exemplar given a value on the other feature. Their results revealed that exemplars tended to share the same feature correlations as exemplars from previously learned categories. Interestingly, this effect was not consistently observed when the given values of a feature fell outside the range of learned values for that particular category. For instance, if the learned Category A exemplars were generally 8 cm to 10 cm in radius and negatively correlated with the angle of the radial line, participants generally produced radial line orientations that were also negatively correlated with the exemplar size when the given size was between 8 cm and 10 cm. However, when the given size fell outside that range for that category (e.g., 14 cm for a Category A exemplar), some participants consistently produced radial line orientations that were not negatively correlated with the size and that also placed the exemplars in a significantly different location in the feature space. These results suggest that when tasked to produce exemplars that fall outside what was previously learned, people show some tendency to generate exemplars that are different (both distributionally and spatially) to what was learned. We explore this idea of contrast more deeply in the following section.

3 And Now for Something (Completely) Different: The Role of Contrast

Although people are *capable* of creating new categories, it is not entirely clear whether (and if so, how) new concepts are systematically made different from what is already known. The hierarchical Bayesian model developed by Jern and Kemp (2013) assumes that differences between observed and generated categories are only due to random variation. The model assumes that generated categories are sampled from the same underlying domain distribution as observed categories, and will thus share a common distributional structure. The model does not make predictions about the *location* of the category within the domain (the perceptual instantiation of category members). Under a strict interpretation of their model, given knowledge of a single category within the domain, the most probable new category to be generated is located in *exactly* the same location and possesses an identical distributional structure. This is not an issue with their model specifically, but using a broader class of standard hierarchical Bayesian models without any additional features (e.g., Griffiths, Sanborn, Canini, & Navarro, 2008; Kemp, Perfors, & Tenenbaum, 2007). Many of these models assume that at some point of the latent generative process the same underlying distribution generates all of the categories and thus, any differences between categories are due to *noise* and should not be *systematic*. If the goal of creating a new category given others in a domain is generating from the posterior predictive distribution then the best a standard hierarchical Bayesian model without a notion of contrast built into the model can do at capturing contrast is to assume that the new category is placed uniformly at random over feature space. But, this defeats the purpose of a hierarchy as it is ignored when determining a new category location![2] Note that this does not mean location information and contrast cannot be captured using hierarchical Bayesian models. In fact, a model instantiating contrast as representativeness is a hierarchical Bayesian model. But, it requires an additional factor, which in our case is having a different goal than predicting exemplars from a new category.

The copy-and-tweak model tested by Jern and Kemp (2013) also claims little about how generated categories should contrast with what is already known. In their simulations, the model was only tested on generation after the learner had been exposed to members of the target category, and so the model's ability to generate a new category from scratch was not evaluated. However, the model's generation is based exclusively on similarity to known members of the *target* category; when there are no members of the target category, generation is presumably random.

To our knowledge, no prior literature has yet explicitly discussed any empirical effects of contrast in category generation. However, some evidence of contrast can be seen in the data from Experiment 3 of Jern and Kemp (2013). We leave a detailed analysis of this data to the supplemental materials.

3.1 Contrast as exemplar dissimilarity: The PACKER Model

As noted above, the constraint that new concepts should differ from what is already known has been largely overlooked in previous work. This is no doubt in part due to the vague definition of what it means for a concept to be "different": A generated category may be different from what is already known in any number of respects. Towards providing a more precise definition of the role of contrast in generation, we formalized contrast in a novel exemplar model, PACKER (*Producing Alike and Contrasting Knowledge using Exemplar Representations*). PACKER explains category generation as a balance between two

[2] It is plausible that some hierarchical Bayesian models could be created that generate categories different from each other. However, this model would not be a standard application or extension of most pre-existing hierarchical Bayesian models. The generative process would need to include a component that presumes contrast, which is precisely the factor of study in this article.

fundamental constraints: The exemplars of the category to be generated should not be similar to known categories, and exemplars within each category should be similar to one another. These ideas are implemented within the well-studied exemplar framework – the PACKER model is an extension of the influential Generalized Context Model of categorization (GCM; Nosofsky, 1984, 1986).

Although, as an exemplar model, one of PACKER's proposals is people represent categories in terms of a collection of stored exemplars, we did not pick it assuming it is the correct model of human categorization. The choice to develop PACKER within an exemplar framework reflects the facts that exemplar models have been thoroughly evaluated, are strongly theoretically motivated, and dominate much of the theoretical and empirical work in categorization. The focus of our work with PACKER concerns how contrast may be captured through the dual constraints of within- and between-class similarity; it is not difficult to imagine how such constraints may be instantiated using alternative frameworks (e.g., Kurtz, 2007; Love, Medin, & Gureckis, 2004; D. J. Smith & Minda, 2000; for a review of categorization models, see Pothos & Wills, 2011).

Both PACKER and the GCM simulate categorization under the assumption that learners represent categories as a collection of exemplars, corresponding to the labeled stimuli they have observed. The exemplars are encoded within a k-dimensional psychological space, and model performance is based on the amount of similarity between the item to be categorized and the stored exemplars. Similarity between two examples, $s\left(x_i, x_j\right)$, is computed as an inverse exponential function of distance (following Attneave, 1950; Shepard, 1957, 1987):

$$s\left(x_i, x_j\right) = \exp\left\{-c\left[\sum_k w_k\left|x_{ik} - x_{jk}\right|^r\right]^{1/r}\right\} \tag{1}$$

where w_k is the attention weighting of dimension k ($w_k \geq 0$ and $\sum_k w_k = 1$), accounting for the relative importance of each dimension in similarity calculations, and c ($c > 0$) is a specificity parameter controlling the spread of exemplar generalization. For simplicity, in our simulations attention will be distributed across each dimension uniformly (unless otherwise noted). The value of r depends on the nature of the experimental conditions being simulated: $r = 1$ is appropriate for separable dimensions, whereas $r = 2$ is appropriate for integral dimensions (e.g., Garner, 1974; Shepard, 1964). In our simulations, we set $r = 1$ due to the separable nature of the stimulus dimensions used in our experiments (see Figure 3).

PACKER (as well as its name) was in part inspired by earlier work from the categorization literature (Hidaka & Smith, 2011; Stewart & Brown, 2005). Hidaka and Smith (2011) argued that natural categories "pack" the values of features such that different categories fill the domain space with distance between one another, while keeping items within the same category close together. Inspired by this idea, PACKER proposes that generation is constrained by both similarity to members of the target category (the category in which a stimulus is being generated) as well as similarity to members of other categories: the most desirable generation candidates are similar to members of the target category and not similar to members of contrast categories. This is achieved by aggregating similarity across known exemplars differently according to class membership. The aggregated similarity $a(y, x)$ between generation candidate y and stored exemplars x is given by:

$$a(y, x) = \sum_j f(x_j) s(y, x_j) \tag{2}$$

where $f(x_j)$ is a function specifying each exemplar's contribution to generation. A negative value for $f(x_j)$ produces a 'repelling' effect (items are less likely to be generated nearby x_j), and a positive value produces an 'attracting' effect (items are more likely to be generated nearby x_j). When $f(x_j) = 0$, the exemplar does not contribute to generation.

PACKER sets $f(x_j)$ depending on exemplar x_j's category membership: $f(x_j) = \theta_t$ if x_j is a member of the target category, and $f(x_j) = -\theta_c$ if x_j is a member of a contrast

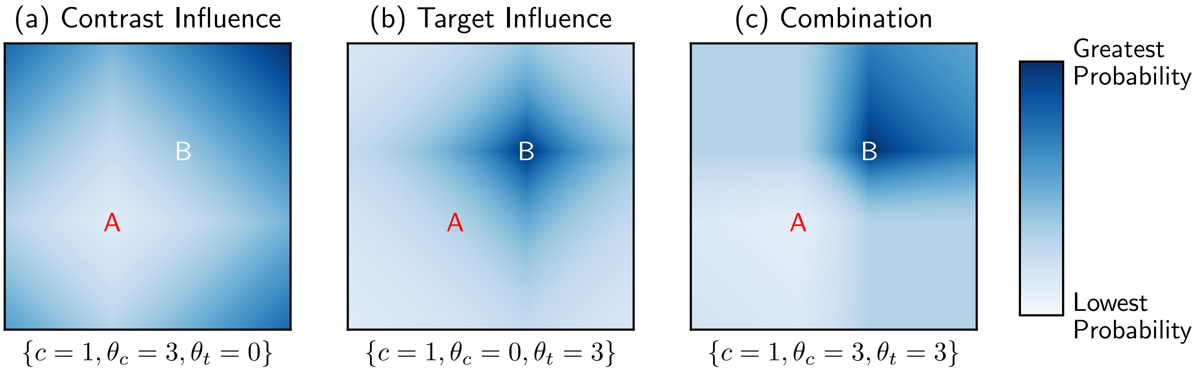

Figure 1: PACKER generation of a category ‘B’ example, following exposure to one member of category ‘A’ and one member of category ‘B’. Predictions are shown for three different parameterizations (differing in values of θ_t and θ_c): *(a)* Predictions based on contrast similarity only. *(b)* Predictions based on target similarity only. *(c)* Predictions with both constraints considered.

category. θ_t and θ_c are free parameters ($0 \leq \theta_t, \theta_c$) controlling the trade-off between within- and between-category similarity. For example, when $\theta_t = \theta_c = 0.5$, $f(x_j) = 0.5$ for members of the target category and $f(x_j) = -0.5$ for members of other categories; thus, the model is likely to generate items that are similar to members of the target category but are not similar to members of other categories. In this way, $\theta_t > 0$ with $\theta_c = 0$ produces exclusive consideration of target-category members, and $\theta_c > 0$ with $\theta_t = 0$ produces exclusive consideration of contrast-category members. The combination of θ_t and θ_c parameters thus specifies a wide breadth of possible approaches; by fitting it to a dataset, one can describe the relative roles of between-category contrast and within-category similarity in generation. See Figure 1 for an illustration of how these parameters control the relative influence of within-category similarity and contrast to other categories when generating a new exemplar.

The probability that a candidate y will be generated is evaluated using an Exponentiated Luce (1977) choice rule. Candidates with greater values of $a(y, x)$ are more likely to be generated than candidates with smaller values:

$$p(y \mid x) = \frac{\exp\{a(y, x)\}}{\sum_i \exp\{a(y_i, x)\}} \tag{3}$$

Note, because θ_t and θ_c are unconstrained, their size plays the role of the response determinism parameter for the exponentiated Luce choice rule.

3.1.1 Relation Between PACKER and Copy-And-Tweak

It is worth noting that PACKER is only one possible exemplar-based account of category generation within our proposed framework. That is, PACKER places specific constraints on the possible values of $f(x_j)$, but other exemplar-based category generation models with drastically different behavior can be formalized in this framework by imposing alternative constraints. For example, PACKER is formally equivalent to the copy-and-tweak model proposed by Jern and Kemp (2013) when $\theta_c = 0$ and $\theta_t > 0$ (in fact, it becomes a continuous-dimension adaptation of the original model). Likewise, when $\theta_t = 0$ and $\theta_c > 0$, PACKER can represent a contrast-only generation mode, relying exclusively on contrast when generating new categories. When $f(x_j) < 0$ for all x_j (regardless of class membership), a "pure-packing" approach is yielded, generating items in unoccupied areas of the domain. Thus, the proposed framework may be used to describe a wide variety of qualitatively distinct generation strategies.

By formalizing a model family where PACKER and copy-and-tweak are different parameterizations of models within the same framework, the comparison between PACKER and copy-and-tweak provides a test of the explanatory value of the contrast mechanism based on exemplar dissimilarity: The account provided by copy-and-tweak will only equal that of PACKER if the contrast mechanism does not offer an advantage (i.e., if $\theta_c > 0$ significantly improves model fits). Note that the purpose of the article is to explore and formally analyze the role of contrast in categorization and thus, we leave extending PACKER to incorporate distributional factors for future work.

3.2 Contrast as representativeness: An extension to Jern and Kemp (2013)

An alternate conceptualization of category contrast is the idea of representativeness – exemplars are generated such that they resemble the target category and are thus more distinct from and less similar to other categories. The general idea of representativeness is not new in categorization. For instance, the literature on the *graded structure* of categories explores how some exemplars are better exemplars of their categories (Barsalou, 1985; Palmeri & Nosofsky, 2001). In this article we adopt the specific formalization of representativeness given by Tenenbaum and Griffiths (2001), where the representativeness $R(y, h)$ of an item y (or in our case, an exemplar) is the relative amount of evidence that is provided by y for a given hypothesis h in a space of hypotheses $\mathcal{H}$ in contrast to all other hypotheses $\mathcal{H}^c = \mathcal{H} \setminus \{h\}$:

$$R(x, h) = \log \frac{p(x|h)}{1 - p(x|\mathcal{H}^c)} = \log \frac{p(x|h)}{\sum_{h' \in \mathcal{H}^c} p(x|h')p(h')}. \tag{4}$$

In Equation 4, $p(h')$ is the prior probability distribution on the hypothesis space that excludes h (i.e., we effectively set $p(h)$ to 0 and then renormalize the priors). Note that when there are only two hypotheses (e.g. two categories) involved in the domain, this prior takes the value of 1 since there is ever only one alternative hypothesis.

The Bayesian formalization of representativeness makes it straightforward to extend the existing hierarchical Bayesian model of category generation developed by Jern and Kemp (2013).[3] Specifically, both models assume that exemplars are sampled from a given category distribution (h in Equation 4). Each category distribution is a multivariate normal parameterized by a location vector μ_k and covariance matrix Σ_k (i.e., $h = \mathcal{N}(\mu_k, \Sigma_k)$). The parameters μ_k and Σ_k are assumed to be samples from a prior

[3] We thank Charles Kemp for the suggestion.

normal-inverse-Wishart (NIW) distribution parameterized by $\mu_0, \Sigma_0, \kappa, \nu$. Mathematically,

$$\mu_k, \Sigma_k | \mu_0, \Sigma_0, \kappa, \nu \sim \text{NIW}(\mu_0, \Sigma_0, \kappa, \nu) \tag{5}$$

$$y | C \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{6}$$

In our simulations, we set the prior mean μ_0 to the center of the feature space and the prior variance to be isotropic ($\Sigma_0 = \lambda\mathbf{I}$ where λ is a free parameter and $\mathbf{I}$ is a d-by-d identity matrix with d representing the number of dimensions or features in the domain). The κ and ν parameters are freely estimated within the constraints $\kappa > d - 1$ and $\nu > 0$.

With the assumption of the NIW prior, the expected location vector μ_k given exemplars observed from category k is:

$$\mu_k = \frac{\kappa\mu_0 + n_k\bar{x}_k}{\kappa + n_k} \tag{7}$$

where n_k is the number of observed exemplars in category k and $\bar{x}_k$ is the observed category mean. Note that if there are no observed exemplars in the target category (i.e., if the model is generating a completely novel category), Equation 7 simplifies to $\mu_k = \mu_0$.

The NIW prior also allows us to infer the category covariance matrix Σ_k by computing the following:

$$\Sigma_k = [\nu\Sigma_D + C_k + \frac{\kappa n_k}{\kappa + n_k}(\bar{x_k} - \mu_k)(\bar{x_k} - \mu_k)^T](\nu + n_k)^{-1} \tag{8}$$

where C_k is the observed category covariance and Σ_D is the domain covariance matrix from which Σ_k samples are obtained. We can infer Σ_D from the observed category covariances C and the prior covariance Σ_0. Specifically,

$$\Sigma_D = \Sigma_0 + \sum_k C_k \tag{9}$$

At this point, both the representativeness model and the hierarchical Bayesian model from Jern and Kemp (2013) are largely identical. However, both models diverge in their computation of exemplar generation probabilities. While the original hierarchical Bayesian model produces novel exemplars with probabilities proportional to the multivariate normal likelihood $p(y|h)$, the representativeness model generates new exemplars according to their representativeness $R(y, h)$ as formulated in Equation 4. In practice, the probability of generating a particular candidate y is obtained using an Exponentiated choice rule (Luce, 1977):

$$p(y) = \frac{\exp(\theta \cdot R(y, h))}{\sum_k \exp(\theta \cdot R(y_k, h_k))}. \tag{10}$$

where θ is a freely estimated response determinism parameter (constrained such that $\theta \geq 0$).

Here, we implement the representativeness mechanism at the level of distributions over category exemplars. However, with a hierarchical model, it is also possible to apply the representativeness mechanism at other levels, such as the level of the distribution of category statistics (i.e., their means and covariances). For simplicity, we focus only on one type of representativeness model (i.e., where the mechanism is implemented at the level of distribution over category exemplars).

Despite both the original hierarchical Bayesian model and our representativeness model sharing identical hierarchical structures, their distinct response processes can yield very different exemplars. Specifically, the hierarchical Bayesian model emphasizes the generation of exemplars that maintain distributional commonalities across categories. In contrast, the representativeness model focuses on the generation of exemplars that are representative of the underlying distribution for a target category. Generally, with the assumption of unimodal underlying distributions, this mechanism of representativeness results in the generation of novel exemplars that are less similar to exemplars from the

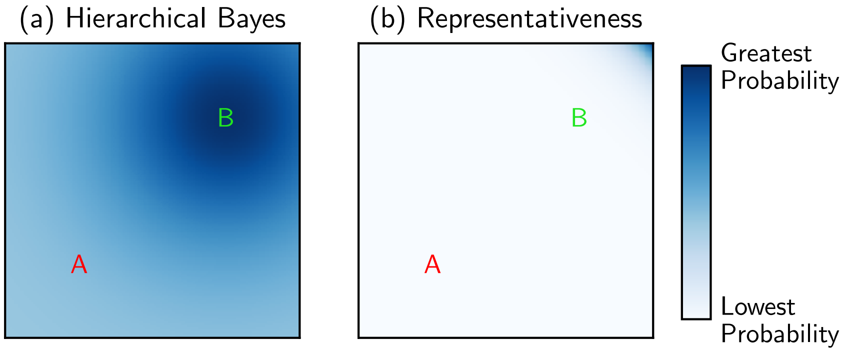

Figure 2: Generation of a category 'B' example, following exposure to one member of category 'A' and one member of category 'B'. Predictions are shown for (a) the original hierarchical Bayesian model and (b) the representativeness model.

known category. This occurs because the representativeness of novel exemplars tend to be highest where the probability density of the known category's underlying distribution is the lowest. We see this illustrated in Figure 2, where the representativeness model displays a strong preference for generating exemplars from target category 'B' that are further away from the contrast category 'A'. Exceptions to this pattern can occur (i.e., the generation of novel exemplars similar to known category exemplars) – we explore this in more detail in Section 7 where we highlight an important condition demonstrating how PACKER and the representativeness model can make qualitatively different predictions.

4 Experiment 1

To begin our investigation, we examined category generation in a well-understood domain using a few disparate category types. We used an artificial stimulus design: A two dimensional domain of squares, varying in color and size (see Figure 3a). These dimensions have been used in numerous classification learning studies (e.g., Conaway & Kurtz, 2016a, 2016b; Nosofsky, Gluck, Palmeri, & McKinley, 1994; Shepard et al., 1961). Unlike those used in the Jern and Kemp (2013) experiments, distance on these physical dimensions aligns more directly with perceptual similarity, allowing us to evaluate the role of contrast in categorization more precisely. It also enables more straightforward comparisons to prior work. We tested the effects of category contrast after learning one category from a set of qualitatively distinct category structures, as shown in Figure 3.

Figures 3b-d show the values of exemplar dimensions belonging to the experimenter-defined categories ('A', or 'Alpha') that participants were assigned to learn about prior to generating a new category. Each participant learned one of the category types during training. In the 'Cluster' type, category A is a tight cluster of examples in the space. Perceptually instantiated, the members of category A might, for example, be large and dark in color. In the 'Row' type, category A has a row pattern across the space, varying along one feature but not the other. Thus, its members might all be dark in color but would vary in size. Finally, in the 'XOR' type, the experimenter-defined category consists of two clusters separated in opposite corners of the space, conforming to the exclusive-or logical structure (e.g., members are small and dark or large and light).

It should be noted that in our experiments the assignment of the perceptual to conceptual dimensions (e.g., $X \rightarrow Size$, $Y \rightarrow Color$) and the direction of variation along each dimension (e.g., $dark \rightarrow light$ or $light \rightarrow dark$) were counterbalanced across participants. The category types in Figure 3 are plotted in conceptual space, rather than perceptual space. Thus, while the conceptual organization of the category types remains constant, each category type may have a different physical instantiation according to the counterbalance assignment. For example, the Cluster type may be large and dark in color, or it may be small and light in color, depending on the assignment and direction of the dimensions. For this reason, below we will discuss generation within a conceptual space, rather than a physically instantiated one.

After learning about an experimenter-defined category, participants are asked to generate examples of a new category. Within this paradigm, an effect of category contrast would be realized if participants prefer to generate items in locations that are distant (i.e., perceptually dissimilar) from members of category A. However, generation is left

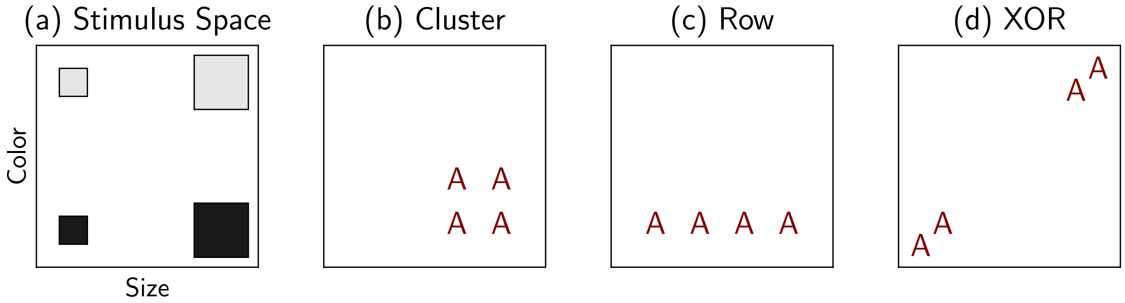

Figure 3: Stimulus domain and category types tested in Experiment 1. Stimuli are not drawn to scale. Dimension and direction assignment (e.g., large to small or small to large) for color and size were counterbalanced over participants.

unconstrained. Critically, participants were not asked to generate something different in the prompt. For example, participants assigned to the Cluster condition may generate a tightly clustered category in the corner opposite of the experimenter-defined category. Alternatively, they may generate a tightly clustered category directly overlapping with the experimenter-defined category. Further, they may even generate an entirely different type of category (e.g., a row category).

Our experimental results also provide a converging test that generated categories tend to share distributional properties with known categories in the domain (Jern & Kemp, 2013; Ward, 1994). From these results, we can predict that, in each condition, participants should generate categories that are distributionally similar to the experimenter-defined category: In the Cluster condition, generated categories should be tightly clustered. In the Row condition, generated categories should vary more along the X-axis than the Y-axis. In XOR condition, generated categories should be widely distributed across both dimensions, and the two dimensions should be positively correlated.

Interestingly, the XOR condition also offers a dissociation between the roles of category contrast and the emulation of distributional structure: widely-distributed, positively-correlated categories would need to lie along the positive diagonal of the space (that is the only place they “fit”), which is already occupied by the experimenter-defined category. Thus, if contrast plays a role, exemplars in the generated categories of participants in the XOR condition may not be positively correlated – they may not be correlated, or perhaps even be negatively correlated. In this case, contrast and statistical regularities would interact, which would be inconsistent with prevailing theories of category generation (Jern & Kemp, 2013).

4.1 Participants and Materials

183 participants were recruited from Amazon Mechanical Turk. Each participant was randomly assigned to one condition: 64 participants were assigned to the Cluster condition, 61 were assigned to the Row condition, and 58 were assigned to the XOR condition (sample sizes differ due to random assignment). Stimuli were squares varying in color (grayscale 9.8%–90.2%) and side length (3.0–5.8cm), see Figure 3. The assignment of perceptual features (color, size) to axes of the conceptual space (X, Y) and the direction of variation along each axis (e.g., *dark* → *light* or *light* → *dark*) were counterbalanced across participants.

4.2 Procedure

As noted in the introduction of this paper, the task of generating members of a new category is well situated as a task in the categorization literature: Whereas classification consists of predicting an object's category label on the basis of its features, inference consists of predicting an observed feature, given a set of observed features and a category label. Generation thus consists of predicting *all* features of an object, given a novel category label. We designed our generation task as an extension of the traditional artificial classification learning paradigm. The task differs from traditional work in creative cognition primarily through the use of an artificial domain, which enables the application of computational models. The use of an artificial domain also requires the addition of a training phase, during which participants learn about the categories in the domain. As a result, unlike most previous studies (e.g., Ward, 1994), participants in our studies have no experience with the domain before the start of the experiment, and the experimenter-defined categories are not hierarchically structured (as are many natural categories).

Participants began the experiment with a short training phase (3 blocks of 4 trials), where they observed exemplars belonging to the 'Alpha' category. Participants were instructed to learn as much as they can about the 'Alpha' category, and that they would answer a series of test questions afterwards. On each trial, a single 'Alpha' category exemplar was presented, and participants were given as much time as they desired to observe it before moving on to the next trial. Each block consisted of a single presentation of each of the members of the 'Alpha' category, in a random order. Participants were shown the range of possible colors and sizes prior to training.

Following the training phase, participants were asked to generate four examples belonging to another category called 'Beta'. As in Jern and Kemp (2013), generation was completed using a sliding-scale interface. Two scales controlled the values of the two dimensions (color, size) for the generated example. An on-screen preview of the example updated whenever one of the features was changed. Participants could generate any example along an evenly-spaced 9x9 grid (including members of the 'Alpha' category), except for any previously generated 'Beta' exemplars. Neither the members of the 'Alpha' category nor the previously generated 'Beta' examples were visible during generation. Prior to beginning the generation phase, participants read the following instructions:

> As it turns out, there is another category of geometric figures called "Beta". Instead of showing you examples of the Beta category, we would like to know what you think is likely to be in the Beta category.
>
> You will now be given the chance to create examples of any size or color in order to show what you expect about the Beta category. You will be asked to produce 4 Beta examples - they can be quite similar or quite different to each other, depending on what you think makes the most sense for the category.

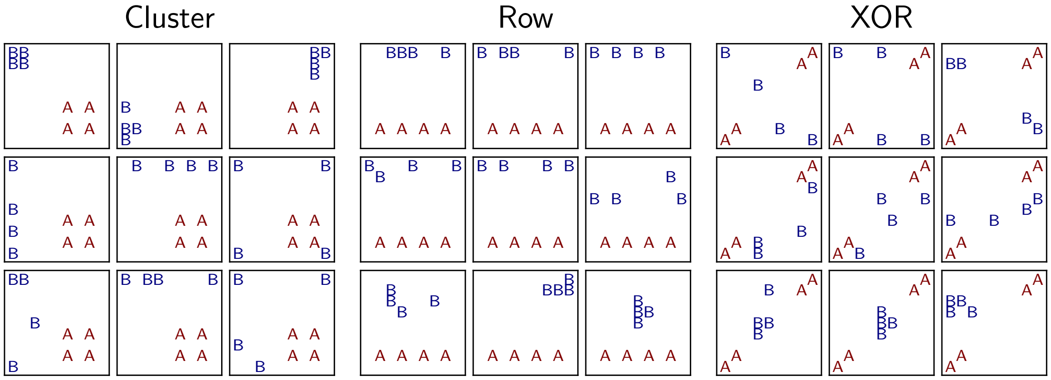

Figure 4: Sample categories generated by participants in Experiment 1. Representative samples from common generation profiles are shown.

Each example needs to be unique, but the computer will let you know if you accidentally create a repeat.

4.3 Results

We observed a substantial degree of individual differences in our data. In Figure 4 we have plotted sample data from several participants, from which it is evident that different participants generated qualitatively different category structures. In this section we will focus on analyzing the data in aggregate, but in later sections we will explore how these individual differences can be explained.

To evaluate the role of contrast, we computed the number of times each stimulus was generated, as a function of its average city-block distance from members of the experimenter-defined “Alpha” category. These data, shown in Figure 5, reveal a clear pattern: Examples that are more distant from members of the experimenter-defined categories are more likely to be generated into a new category. This supports the notion that contrast is a fundamental constraint on how categories are related to one another and that statistical regularity alone is insufficient.

Figure 5 also depicts, for each participant, the average distance of members within

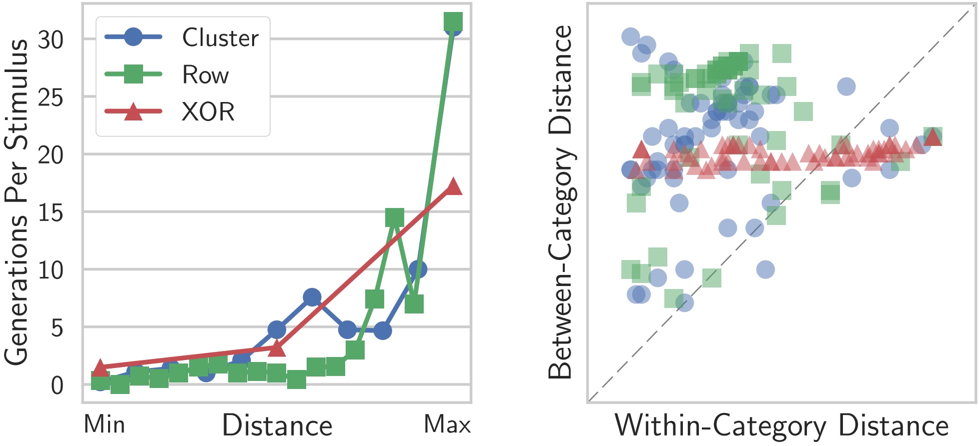

Figure 5: Experiment 1 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

the generated category (*within-category* distance) against the average distance between members of the generated and experimenter-defined category (*between-category* distance). The narrow distribution of between-category distances in the XOR condition reflects the widely distributed nature of the experimenter-defined category, reducing the possible distances to members of the participant-generated category. These data reveal a systematic pattern: The majority of participants generated categories with greater between-category distance than within-category distance. That is, members of the generated category tended to be more similar to one another than to members of the experimenter-defined category. To evaluate this claim quantitatively, we conducted t-tests comparing the amount of within- and between- class distance in each condition. All conditions possessed greater between-category distance: Cluster, $t(63) = 11.43$, $p < .001$; Row, $t(60) = 13.16$, $p < .001$; and XOR, $t(57) = 3.64$, $p < .001$. These results provide further evidence of an effect of category contrast: Participants prefer to generate categories that are dissimilar to the learned category but maintain some level of internal cohesion.

A secondary goal of this experiment was to examine whether we replicate the classic result that generated categories often possess the same distributional properties as previously-known categories. Given the increased importance of replication within psychology (Zwaan, Etz, Lucas, & Donnellan, 2018), it is important as it serves as a conceptual replication of Jern and Kemp (2013). For each generated category, we computed the category range along each axis (X, Y), as well as the correlation between features. These data, shown in Figure 6, reveal broad individual differences: Within each condition, some participants generated categories spanning the entire X- and Y- axis whereas other participants categories that spanned very little along each. Likewise, in each condition participants generated categories possessing strongly positive, neutral, and strongly negative correlations between the dimensions. Comparing the distributional statistics between conditions yields a broad yet, as we will see, misleading replication of the classic effect.

With respect to ranges along each axis (X, Y), the generated categories from each condition tend to reflect the ranges of the experimenter-defined categories. The categories generated in the Cluster condition were less widely distributed along the X-axis compared to Row, $t(123) = 5.61$, $p < .001$, and XOR, $t(120) = 2.68$, $p < .01$. Categories generated in the XOR condition were also less widely distributed along the X-axis compared to Row, $t(117) = 2.56$, $p = .046$. This latter effect was not expected because the experimenter-defined categories for XOR and Row had similar X-ranges. However, the key finding is that categories from the Cluster condition tended to be more tightly distributed along the X-axis.

Likewise, categories generated in the Row condition had less Y-axis range compared to Cluster, $t(123) = 4.57$, $p < .001$ and XOR, $t(117) = 9.26$, $p < .001$, and categories from the Cluster condition had less Y-axis range compared to XOR, $t(120) = 3.95$, $p < .001$. As expected, the correlations in the Cluster and Row conditions were not systematically positive or negative ($ps > .1$). However, the generated categories in the XOR condition tended to possess *negatively* correlated dimensions, $t(57) = 2.04$, $p = .046$. This finding is notable, as it is the opposite of what would be expected based on previous literature (Jern

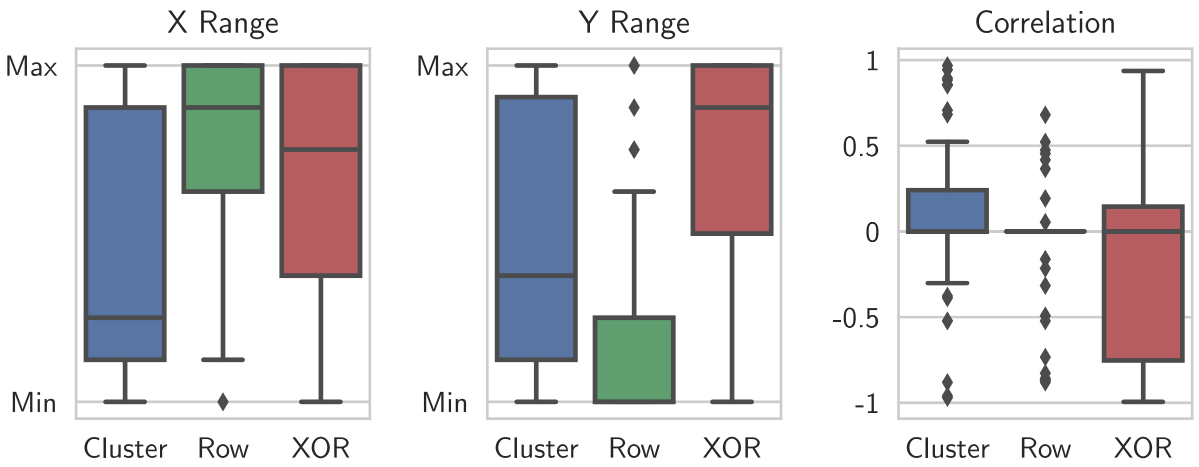

Figure 6: Box-plots of the distributional statistics from the categories generated in Experiment 1. Boxes depict the median and quartiles of each condition, with whiskers placed at 1.5 IQR. All points outside this region are marked individually.

& Kemp, 2013), assuming learners are emulating the distributional structure of the experimenter-defined class (which possesses perfectly positively correlated features).

We believe that the failure to replicate the emulation of dimension correlation is because participants in Jern and Kemp (2013) could differentiate the generated category on a third dimension (hue) to maintain the statistical regularities on the other two dimensions. In addition, in our XOR condition the stimuli were constrained to the corners of the feature space, whereas in the positive diagonal condition of Jern and Kemp (2013) the stimuli had no such constraint. Although the correlation in the XOR condition is significantly negative, it is clear from the box-plot in Figure 6 that it would be inappropriate to make a strong conclusion (e.g., the median is close to zero). However, we can conclude with confidence that there are situations where people do not emulate the distributional structure of the given category. This indicates that there is more to category generation than the emulation of distributional structure of other categories in the domain. Further, as we will discuss in more detail in the model-based analysis section, this is expected by our proposal that contrast is a fundamental principle in categorization.

4.4 Discussion

In Experiment 1 we evaluated the influence of category contrast on category generation, given qualitatively different types of categories. We found strong evidence for effects of category contrast in each condition: Participants were more likely to generate stimuli that are more distant from (i.e., less similar to) members of a previously-learned category, and members of participant-generated categories tended to be more similar to one another than to members of previously-learned categories. We also partially replicated the classic finding that the distributional structure of generated categories reflects that of previously learned categories (Jern & Kemp, 2013; Ward, 1994): Members of generated categories were more widely distributed along dimensions which were widely distributed in the experimenter-defined category.

Notably, however, we also found that participants who learned an XOR category (composed of exemplars following a positive diagonal, see Figure 3) tended to generate items according to a *negative* feature correlation – the opposite of what was present in the previously learned category. While this may be difficult to account for under existing theoretical approaches (which assume generated categories follow the same distributional structure as known categories), it can be concisely explained from a category contrast perspective. Specifically, within the XOR condition, individuals who seek to generate a category that is perceptually distinct from what is already known are left with only the upper-left and bottom-right quadrants of the space, as members of the previously-learned XOR category lie in the bottom-left and top-right. If examples are generated into both of the available quadrants, the generated category will possess a strongly negative correlation, opposing that of the experimenter-defined class.

Thus, the core results of Experiment 1 indicate that generated categories can systematically differ from what we would expect based on prior work. The negative (or null) correlations observed in the XOR condition suggests an interesting interaction between contrast with a given category and emulation of statistical properties. That is, the

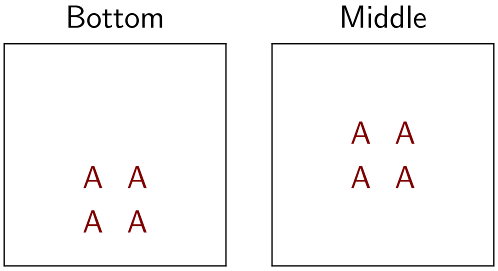

Figure 7: Category types tested in Experiment 2.

constraints on generation imposed by category contrast may not simply influence the *location* of generated categories, but also their distributional structure. In Experiment 2, we test this claim more systematically.

5 Experiment 2

To test whether category contrast influences the distributional structure of generated categories, we sought to identify conditions in which differences in the distributional structure of generated categories cannot be explained by the distributional structure of the experimenter-defined category. We created two new category types (depicted in Figure 7) that possess identical distributional structures (both are tight clusters of examples with no correlation between features), as they only differ in their Y-axis position: the 'Bottom' category lies near the bottom of the space, and the 'Middle' category lies in the center. The distributional equality of these conditions is key to the design of the experiment: If the distributional structure of previously learned categories were the only influence on the generated categories, we should observe no difference in the categories participants generate between these two conditions. Will participants distribute their generated category differently between conditions due to the differences in the available empty feature space for generating a new category?

If category contrast influences the distributional structure of the categories people generate, then we should observe different types of categories according to the shape of the space that is *unoccupied* by members of previously learned categories. The difference in the Y-axis position between the Bottom and Middle conditions produces a considerable change to the shape of the unoccupied space. Participants assigned to learn the Bottom category should be less likely to generate exemplars into the lower regions of the feature space (as these areas possess greater similarity to members of the Bottom category), preferring instead to distribute exemplars across the upper region of the space. This constraint is lifted in the Middle condition, as the Middle category exemplars are equidistant to the upper and lower regions of the space. Accordingly, participants should be more likely to utilize both of these areas. Thus, if category contrast influences the distributional structure of generated categories, we should observe more participants in the Middle condition that generate examples above *and* below the experimenter-defined category.

5.1 Participants, Materials, and Procedure

122 participants were recruited from Amazon Mechanical Turk. 61 participants were randomly assigned to the Middle and Bottom conditions each. The stimulus space and procedure were exactly as in Experiment 1. Participants first completed a short training phase, followed by the generation phase. The only difference from Experiment 1 was the category types given to participants.

5.2 Results

As in Experiment 1, we observed broad differences in the generation approach taken by different participants. To characterize the nature of these differences, Figure 8 depicts sample categories generated by participants. The data from each condition are organized into four columns based on commonly observed patterns of generation: a ‘Cluster’ type of tightly-clustered examples, ‘Row’ and ‘Column’ types of exemplars widely distributed along the one axis but narrowly along the other, and a ‘Corners’ type, wherein participants placed exemplars in disparate corners of the space. As before, in this section we focus on

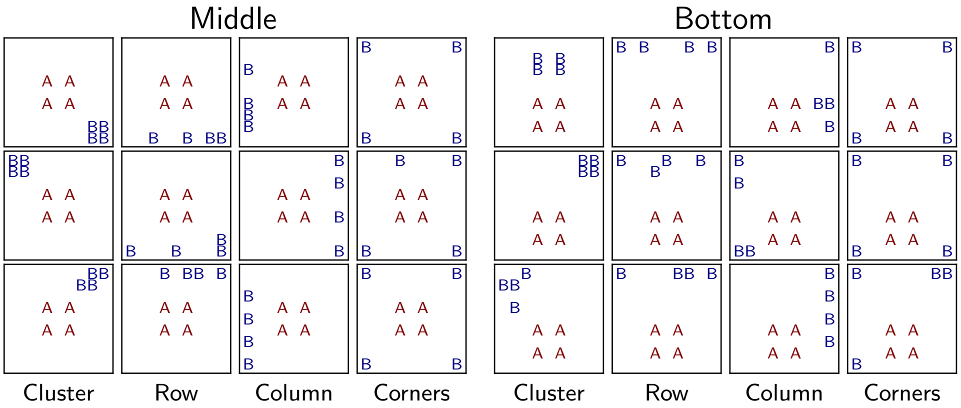

Figure 8: Sample categories generated in Experiment 2.

analyzing the data in aggregate, but in later sections we will focus more specifically on explaining the individual differences.

We began our analysis by testing for the broad influence of category contrast on generation. As in Experiment 1, we computed the frequency each stimulus was generated as a function of its average distance from members of the experimenter-defined category, as well as each participant's average within- and between- category distance. These data, shown in Figure 9, yield very similar results. Participants generated stimuli that are distant from members of the experimenter-defined category, and the categories in each condition tended to possess more between-category than within-category distance: Bottom, $t(60) = 5.5$, $p < .001$; Middle, $t(60) = 2.71$, $p < .01$. We did, however, observe a notable subgroup of participants in each condition who generated categories with more within-category than between-category distance. Upon manual inspection, many of these individuals appear to have assumed a 'Corners' strategy, placing exemplars in disparate corners of the space, thus producing much more within-category distance, see Figure 8 for examples.

To explore the distributional structure of the generated categories, we computed the range of exemplars along each axis (X, Y), as well as the correlation between features. These data, shown in Figure 10, again demonstrate the degree of individual differences

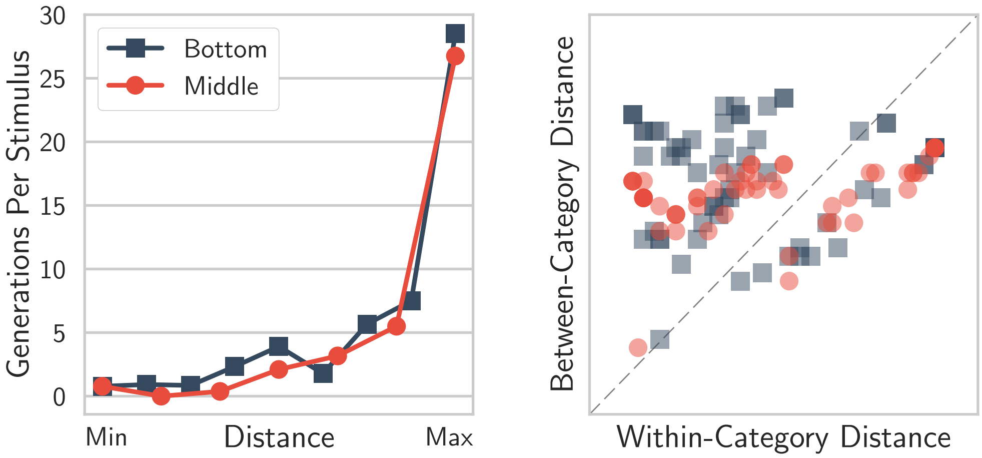

Figure 9: Experiment 2 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

observed in our study. In each condition, we observed tightly clustered and widely distributed categories along each dimension. Although most participants generated uncorrelated categories in both conditions, many still produced positively and negatively correlated categories.

As noted above, if the distributional structure of generated categories is influenced by the shape of the space not occupied by members of known categories, then participants in the Middle condition would be more likely to place exemplars in the upper *and* lower regions of the space, as members of the experimenter-defined category are equidistant from these regions. Participants in the Bottom condition should be less likely to generate category members in the bottom regions because members of the experimenter-defined category are located there. One way to test these predictions is to analyze the Y-axis ranges of the generated categories: If Middle participants utilize the upper and lower regions of the space, their categories should vary more along the Y-axis. T-Tests comparing the conditions on the distributional statistics, however, reveal few between-group differences: the conditions do not differ with respect to X-axis range, Y-axis range, or feature correlations ($ps > 0.17$).

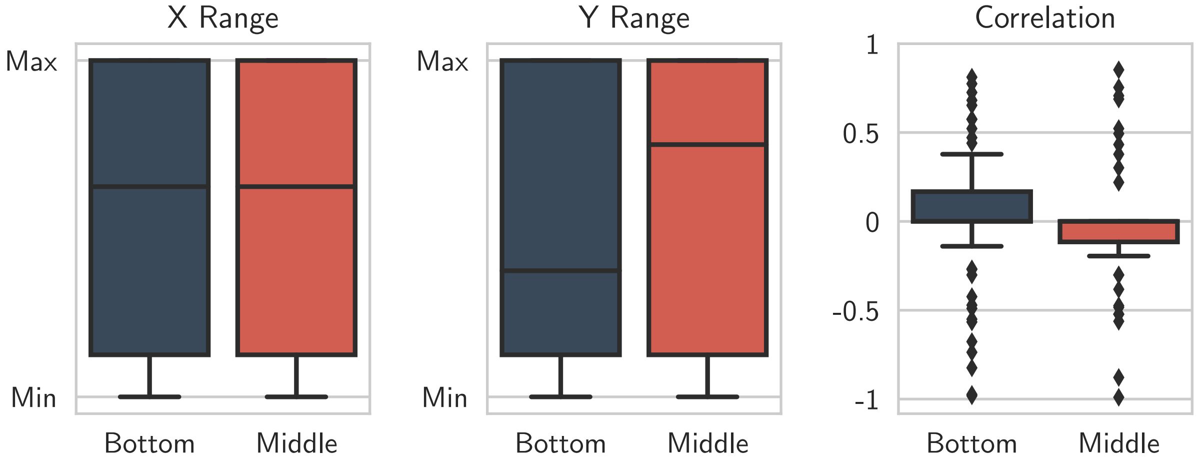

Figure 10: Box-plots of the distributional statistics from the categories generated in Experiment 2. Boxes depict the median and quartiles of each condition, with whiskers placed at 1.5 IQR. All points outside this region are marked individually.

However, our ability to detect differences in Y-axis range using a standard *t*-test between the conditions is, in this case, diminished due to the non-normality of the data (Shapiro-Wilk normality test $W = 0.77, p < .001$ for the Middle condition and $W = 0.85, p < .001$ for the Bottom condition). Figure 11 depicts the Y-axis range and Y-axis position of exemplars generated by each participant. The categories are sorted by overall range, and then colored by training condition. These data reveal that there were nearly as many participants who generated categories spanning the entire Y-axis as those who generated categories spanning almost none of the Y-axis. Indeed, the median produced Y-axis range is much smaller in the Bottom than Middle condition, whereas they are essentially identical in X-axis range. The non-normality of the Y-axis range distributions thus requires that we use a different analytic approach to addressing the experiment's main question.

Because our main prediction concerns the generation of exemplars within the upper and lower regions of the domain, we compared the conditions in terms of the frequency with which participants generated examples above and below the categories. Specifically, we counted the number of participants in each condition who placed at least one 'Beta'

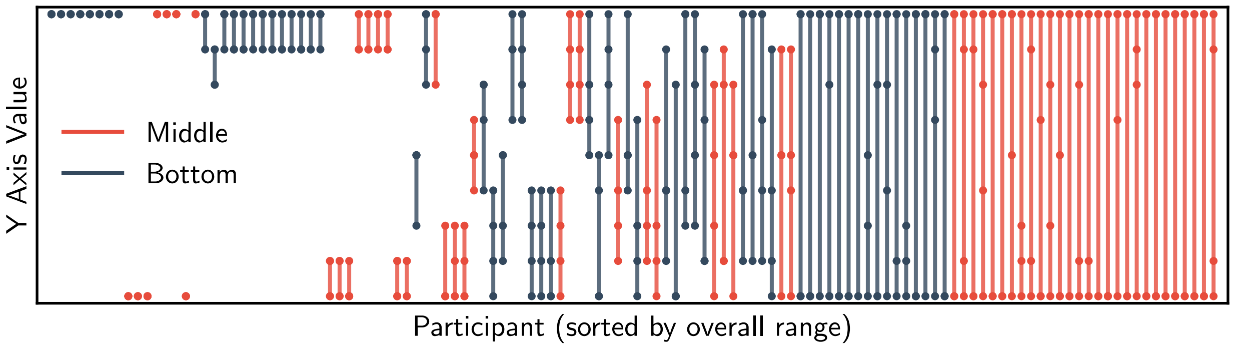

Figure 11: Y-axis range and position of the participant-generated categories from Experiment 2. Each line corresponds to a participant's category, with notches corresponding to the Y-axis position of exemplars within the category (notches may overlap). Participants are sorted by overall range, and then by condition.

exemplar on the top and bottom 'rows' of the space (the maximum and minimum possible Y-axis value, respectively). The resulting contingencies data are shown in Table 1.

Firstly, it should be noted that nearly every participant utilized the top and/or bottom rows: only 10/122 participants generated their category entirely within the interior region. Fisher's Exact Tests comparing the conditions reveal that more Middle participants generated an exemplar in the bottom row, $p < .001$, again demonstrating the role of contrast in guiding where exemplars are generated. The conditions did not differ in use of the top of the space, $p = .16$, however, more Middle participants placed exemplars in the top *and* bottom rows, $p = .04$. The latter effect is of particular interest here, as it indicates that the shape of the unoccupied space exerts some influence on the distributional structure of generated categories: Participants in the Middle condition were more likely to generate a category spanning the entire Y-axis. Thus, the distributional structure of the generated categories can be influenced without any change to the distributional structure of the given category. Rather, it can be affected by category contrast alone.

Table 1: Experiment 2 results.

Middle	Used top row	No top row
Used bottom row	28	18
No bottom row	11	4

Bottom	Used top row	No top row
Used bottom row	16	8
No bottom row	31	6

5.3 Discussion

In Experiment 2, we replicated the core findings from Experiment 1. Stimuli are more likely to be generated if they are distant from exemplars in other categories, and most participants generate categories with more between-category than within-category distance. However, we additionally found that the *position* of a previously learned category (rather than its distributional structure) influences the types of categories people generate: Participants who learned the 'Middle' type were more likely to generate categories spanning the entire Y-axis of the space. Participants who learned the 'Bottom' type were less likely to do so as a result of the presence of opposite category exemplars in the lower regions of the space.

This finding cannot be explained from the perspective that the distributional structure of previously learned categories is the sole determinant of the distributional structure of generated categories. However, the observed behavior is expected from a category contrast perspective: Participants seeking to generate a perceptually distinct category will be more likely to use areas of space that are unoccupied by exemplars belonging to previously learned categories. In the Middle condition, the upper and lower regions of space are equidistant from members of the experimenter-defined category, whereas in the Bottom condition, the lower region of the space is closer to members of the experimenter-defined category. Thus, while Middle participants may form categories around the use of the equally unoccupied areas, the same is not true for the Bottom condition.

6 Model-based Analyses of Experiments 1 and 2

Experiments 1 and 2 revealed systematic and strong effects of category contrast on category generation. In this section, we analyze the performance of the different formal models at explaining the experimental results. Specifically, we present simulations from the two novel contrast models: PACKER and the representativeness model, and compare them to two models that do not incorporate contrast: the copy-and-tweak model (discussed in Section 3.1.1) and an implementation of the hierarchical Bayesian model proposed by Jern and Kemp (2013), described in-depth in Appendix A. For our simulations here, the copy-and-tweak model is defined as a variant of PACKER with the θ_c parameter constrained to be zero. The comparison of this set of models serves to highlight the explanatory role of contrast in categorization: If contrast affords little explanatory advantage, then there should be few differences in performance between PACKER and copy-and-tweak, or between the representativeness and hierarchical Bayesian model. The comparison between these models can also emphasize the necessity of contrast and demonstrate that generation cannot be explained entirely through the emulation of distributional structure. Each model has complementary strengths and weaknesses: Whereas PACKER and copy-and-tweak are relatively insensitive to the distributional structure of learned categories (relying only on exemplar similarities), the representativeness and hierarchical Bayesian model generates categories exclusively on the basis of knowledge of how existing classes are distributed.

Our approach in this section is to first broadly evaluate and compare the quality of each model's account to our entire dataset (Experiments 1 and 2 combined), then analyze the ability for each model to explain individual differences in each experiment, and lastly we describe the strengths and weakness of each model's account of category generation.

Table 2: Results of model-fitting to the combined datasets from Experiments 1 and 2. Note that smaller AIC values correspond to better model fits (adjusted for number of parameters)

PACKER	**Copy & Tweak**	**Representativeness**	**Hierarchical Bayesian**
$AIC = 9069$	$AIC = 9813$	$AIC = 8783$	$AIC = 9881$
$L = -4531$	$L = -4905$	$L = -4388$	$L = -4937$
$c = 0.51$	$c = 3.22$	$\kappa = 12.23$	$\kappa < 0.001$
$\theta_c = 3.09$	$\theta_c = 0$ (fixed)	$\nu = 1.00$	$\nu = 5.44$
$\theta_t = 3.47$	$\theta_t = 3.00$	$\lambda = 7.04$	$\lambda = 0.06$
		$\theta = 10.21$	$\theta = 3.09$

6.1 Parameter-Fitting

To obtain a global measure of the quality of each model's account, we fit the parameters of each model to our entire dataset (Experiments 1 and 2 combined), using a hill-climbing algorithm which maximized the log-likelihood of the model's predictions of the observed responses (1220 responses from 305 total participants). We fit three parameters in the PACKER model (c, θ_t, and θ_c; see Section 3.1), as well as four in the representativeness model and the hierarchical Bayesian model (κ, λ, ν, and θ; see Section 3.2 and Appendix A respectively). We fit only two parameters for the copy-and-tweak model (c, and θ_t), as θ_c is held constant ($\theta_c = 0$). Attention (w, see Equation 1) in PACKER and copy-and-tweak was set uniformly. Parameters were not allowed to vary between participants or conditions – the goal was to obtain the best-fitting values to our entire dataset.

Table 2 contains the model fits. Due to the uneven number of fitted parameters among the models, we compare the model fits using the Akaike Information Criterion (AIC; Akaike, 1974), where smaller values correspond to better fits (discounted by model complexity as measured by the number of parameters). The same qualitative results were obtained with alternative model comparison metrics (e.g., BIC, Schwarz, 1978; AIC_C, Hurvich & Tsai, 1989). In addition to AIC, Table 2 contains the corresponding log-likelihood (L) and the best-fitting parameter values. These results reveal strong model differentiation: both contrast models (PACKER and the representativeness model) achieved far better fits compared to their non-contrast counterparts: copy-and-tweak and the hierarchical Bayesian model respectively. Interestingly between the contrast models, the hierarchical model (representativeness) outperformed the exemplar-based theory (PACKER), whereas between the non-contrast models, the reverse is observed. Specifically, here the exemplar-based model (copy-and-tweak) performed somewhat better than the hierarchical Bayesian model.

While PACKER's advantage over copy-and-tweak may tentatively be attributed to the model's sensitivity to category contrast (this will be explored in detail below), the advantage shown by copy-and-tweak over the hierarchical Bayesian model may be attributed to its exemplar-based representation of category B, as opposed to forcing a prototype-based representation as assumed by the hierarchical Bayesian model. As observed in Figures 4 and 8, the generated categories we observed were often widely distributed, with no items near the category prototype. This aspect of the data is inconsistent with the multivariate normal distributions (similar to prototypes) used to represent categories in the Jern and Kemp (2013) model, but can be easily accounted for using an exemplar-based approach. Interestingly, representativeness using a prototype approach fits better than an exemplar-based approach.

A key distinction between the contrast and classical models is that only the contrast models are capable of making strong predictions about the location of new category members when the target class is entirely novel (i.e., no member of the category has been observed). Under these circumstances, there are no examples to copy, and thus the copy-and-tweak model predicts that items are generated at random. Likewise, with no observations on which to condition the category distribution, the hierarchical Bayesian model also picks an item at random.

Thus, it is possible that the failure of the classical models is simply due to their inability to explain each participant's first trial (generating the first item in the 'Beta' category). To ensure this is not driving our results, we conducted an identical set of simulations as above, excluding the first trial (leaving 915 responses in the dataset): Again, the representativeness model ($L = -3286$, $AIC = 6580$) and PACKER ($L = -3377$, $AIC = 6759$) achieved better fits than the copy-and-tweak ($L = -3564$, $AIC = 7132$) and hierarchical Bayesian ($L = -3597$, $AIC = 7201$) models.

Finally, because copy-and-tweak is nested within PACKER, we can use a likelihood ratio test to compare the two models. PACKER explains the aggregate data significantly better than copy-and-tweak ($\chi^2(1) = 747, p < .001$ for all data and $\chi^2(1) = 375, p < .001$ excluding the first example), providing further evidence that category generation is better explained when contrast is considered.

Through comparison with the copy-and-tweak model, Figure 12 more clearly demonstrates the robustness of the explanatory gains yielded by PACKER's category contrast mechanism. It displays the log-likelihood of the participants' results under PACKER as a function of the θ_c parameter. The model's other parameters (c, θ_t) were set according to copy-and-tweak's best fits from Table 2, and thus when $\theta_c = 0$, the models are equivalent. The figure clearly shows a "sweet spot": a convex region in which PACKER achieves superior fits as a result of changes to θ_c. The best fitting values lie well above the value of 0 assumed by the copy-and-tweak model, which demonstrates the robustness of the contrast effect (though note PACKER achieves even better fits when its parameters are fitted together, as in Table 2). In sum, the data are better explained when both within-category similarity and category contrast is considered.

6.2 Individual Differences

As noted in Experiments 1 and 2, we observed a great deal of individual differences in the types of categories that participants generated. Within each condition, there were a wide variety of category types, such as row and column categories (see Figures 4 and 8). The simulations reported above serve to evaluate the models while considering the entire dataset, but a secondary goal of any formal account should be to provide some explanation of how different profiles of performance emerge. Many of the individual generation profiles

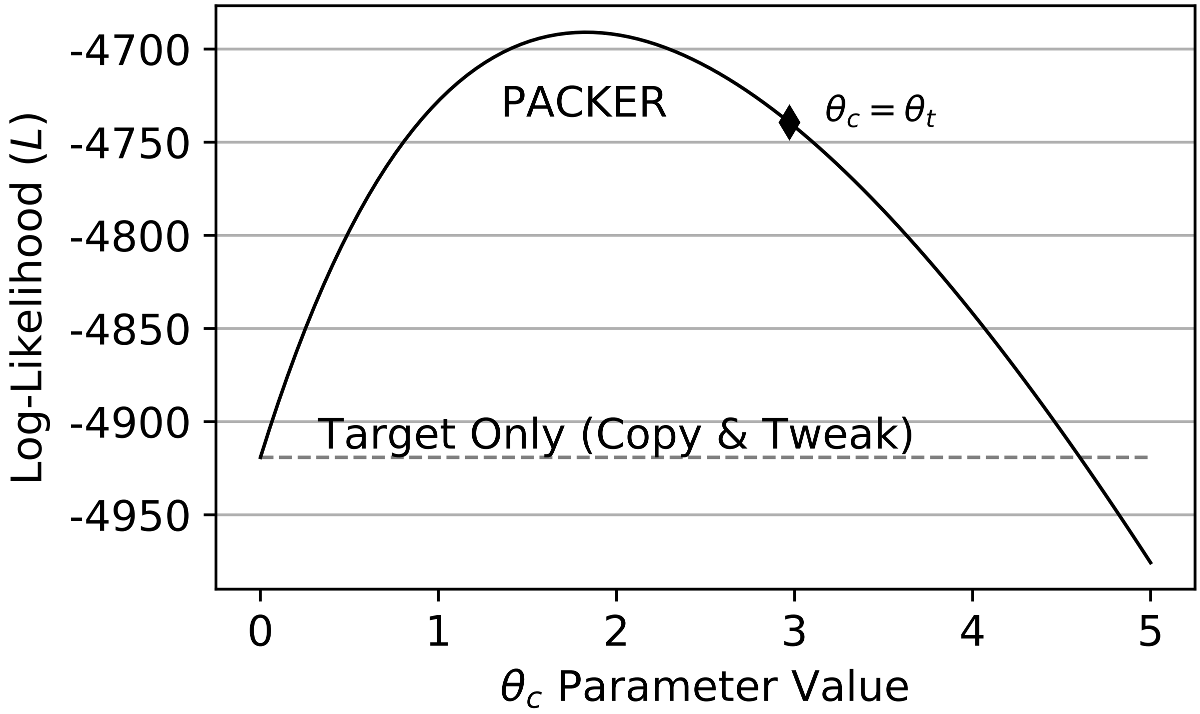

Figure 12: PACKER's fit as a function of its prioritization of within-category and between-category similarity (using the θ_c parameter) . To facilitate comparison, PACKER's other parameters (c, θ_t) were set to the best fitting values obtained for copy-and-tweak in Table 2. The black diamond marker indicates the log-likelihood for the point where $\theta_c = \theta_t$.

we observed can be described with the models simply by tuning the model's parameters in a principled manner. In this section, we describe more specifically how the most frequently observed profiles can be realized.

By manual inspection, it is evident that the most common profiles of generation consist of: (A) a tightly-distributed 'cluster' of examples, (B) 'row'- and 'column'-like arrangements (varying widely along one dimension but not the other), and (C) a 'corners' arrangement with examples placed into disparate corners of the space. These four profiles are distinct in terms of the distribution of the generated category along each dimension: Whereas the cluster profile is tightly distributed along both dimensions, the row and column profiles are tightly distributed along just one dimension. Finally, the corners profile is widely distributed along both dimensions.

In the framework proposed by PACKER, the cluster and corners profiles arise based on different prioritization of within-category similarity versus between-category contrast, and the row and column profiles arise based on the prioritization of each dimension in the computation of similarity. For example, in the cluster profile, there is a high degree of within-category similarity along both dimensions, whereas in the corners profile there is minimal within-category similarity. Thus, PACKER's proposal is that these individual differences arise as a result of different priorities: While the tight cluster configuration can be considered PACKER's 'default' mode (as it maximizes within-category similarity), the corners profile can be produced when between-category contrast is put at a higher priority (i.e., $\theta_c > \theta_t$).

Likewise, in the row and column profiles, there is a large degree of within-category similarity along one dimension but not the other. These differences likely arise due to a differential focus on one dimension over another, and thus they can be produced by changes to PACKER's attention weights, w_1 and w_2 (see Equation 1). Traditionally, the attention weights in exemplar models are thought to reflect the diagnostic value of each dimension towards classifying the known category members (Kruschke, 1992; Nosofsky,

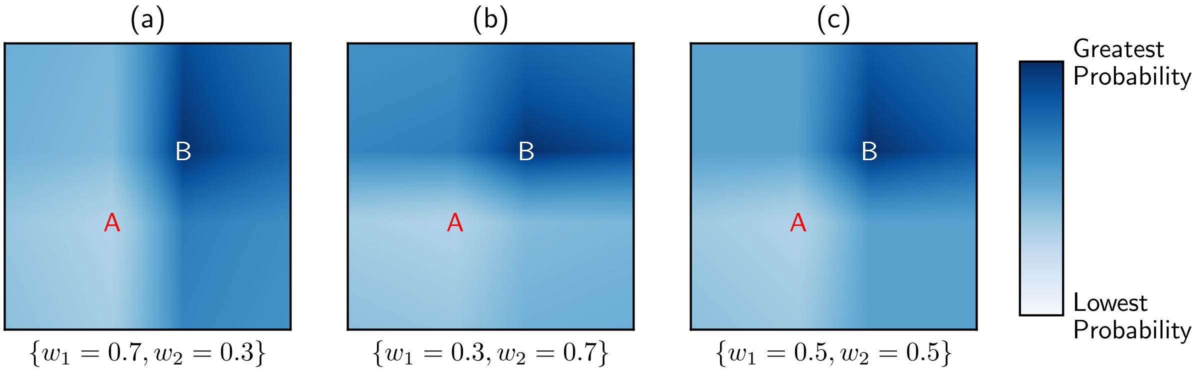

Figure 13: PACKER generation of a category 'B' example, following exposure to one member of category 'A' and one member of category 'B'. Predictions are shown for different attention settings: *(a)* Increased weighting of the X-axis. *(b)* Increased weighting of the Y-axis. *(c)* Uniform weighting (identical to Figure 1).

1984, 1986), but within a generation context the weights specify the importance of within- and between-category similarity along each dimension. For example, if all of attention is allocated along the X-axis ($w_1 = 1$ and $w_2 = 0$), similarity along the Y-axis no longer influences performance. As a result, PACKER will create categories that are more widely distributed along the Y-axis, as similarity is not taken into account along that dimension. As a general principle, differentially weighting one dimension will result in the generation of categories that are more widely distributed along the ignored dimension, conforming to a row- or column-like arrangement. See Figure 13 for a depiction of how attention influences PACKER's performance.

As in PACKER, changes in the parameter settings of the copy-and-tweak model can also be used to produce different patterns of generation. Indeed, as copy-and-tweak is simply a special case of the PACKER model, the attention weights operate exactly as described above to produce row- and column-like categories. However, because the model is not influenced by category contrast, it is biased toward generating tightly clustered categories, as new items are always most likely to be generated near known examples of the target category. Thus, the lack of a contrast mechanism prevents the model from explaining why some individuals widely distribute their categories to the corners of the space.

For both the hierarchical Bayesian and the representativeness models, the prior domain covariance matrix Σ_0 can be used to explain the generation of row-like and column-like categories. This covariance matrix specifies the amount of variance assumed along each dimension (as well as the correlations between dimensions) across the domain of categories. The covariance matrix for a newly generated category, Σ_B, is based on the assumed Σ_0 as well as the distributions of previously learned categories (see Appendix A). Thus, the importance of each feature can be coded into Σ_0 to alter the dimensional variance of generated categories. Because the hierarchical Bayesian model possesses no mechanism to account for category contrast, the model is most likely to generate new items that are similar to known examples of the target category with no regard for how different it is to the contrast category. However, the representativeness model predicts that new exemplars should provide more relative evidence to the 'Beta' category, accounting for the tendency of row-like and column-like profiles occupying the edge of the feature space.

While the copy-and-tweak and hierarchical Bayesian models possess mechanisms to explain row- and column-like categories, they cannot easily explain why some individuals widely distribute their generated categories into disparate corners of the space. This, however, reveals a more general limitation: According to the copy-and-tweak and hierarchical Bayesian models, the distributional structure of generated categories is *independent* of their location within the domain. For example, although the copy-and-tweak or hierarchical Bayesian models can be parameterized to generate row- or column-like categories, there is no mechanism in place to ensure that what is generated will be distinct from what is already known. In the next subsection, we explore this prediction through an analysis of the interdependence between distributional structure and location in category generation.

6.3 Category Location vs. Distributional Structure

As noted above, while all three models make clear claims about the internal structure of generated categories, the copy-and-tweak and hierarchical Bayesian models do not make any claims about how generated categories should differ from what is already known. However, as we observed in the results of Experiment 2, the distributional structure of a category is not always independent of its location within the domain. To demonstrate this point in more depth, we computed the X- and Y- axis ranges of every participant-generated category. Taking the difference between these values $(X - Y)$ produces a measure of each category's orientation in the space: positive difference scores correspond to categories with more X-axis range (horizontally aligned, 'Row' categories), whereas negative difference scores indicate the opposite (vertically aligned, 'Column' categories). Neutral differences scores indicate there was an equal amount of X- and Y-axis range, which can be produced by a number of different category types ('Clusters', 'Corners', etc; see Figures 4 and 8). By plotting, for each possible stimulus, the difference scores of categories it was generated within, we can relate the distributional structure of generated categories to their location within the domain.

However, because many stimuli were infrequently generated (such items near members of the 'Alpha' category), we cannot simply compute the empirical average of the difference scores, as infrequently generated stimuli would be likely to show artificially strong differences. Instead, we used a Bayesian analysis to estimate the mean μ_x on the assumption that the scores x for each stimulus are normally distributed with an unknown mean and unknown standard deviation. The conjugate Normal-Inverse Gamma distribution provides a straightforward method for this estimation:

$$\mu_x = \frac{\nu_0\mu_0 + \sum x}{\nu_0 + n} \tag{11}$$

where μ_0 is the prior mean, ν_0 is a prior scale parameter (controlling the weighting of the

μ_0), and n is the number of categories in which the stimulus was a member (i.e., the number of scores in x). The default assumption is that there is an equal amount of range along the X- and Y-axes, and so we set $\mu_0 = 0$. Likewise, to give a moderate amount of weighting to the prior mean we set $\nu_0 = 1$, though the results are robust to a range of values. Within this approach, the resulting aggregation is a trade-off between the number of generations and the strength of the range difference within each generated category. Infrequently generated stimuli, as well as those with mixed positive and negative scores, are given neutral difference scores.

The results of our analysis are shown in Figure 14 for the experiment and model results[4]. The left-most column of Figure 14 displays the effect of category location and contrast on the distributional structure of the category generated by participants. These data reveal strong and consistent patterns across all the conditions we tested in Experiments 1 and 2: Generated categories are more tightly distributed along the axis in which they are distinct. For example, in the 'Cluster' condition, exemplars in the bottom-left of the space are more often generated into vertically aligned categories, and exemplars in the top-right are more often generated into horizontally aligned categories. Similarly, in the 'Bottom' and 'Middle' conditions, horizontally aligned categories are generated above and below the experimenter-defined categories, while vertically-aligned categories are generated to the sides. In the 'Row' condition, most categories are horizontally aligned, and lie along the upper areas of the space. There are no strong range difference patterns in the XOR condition.

These patterns of performance clearly depict the interdependence between the distributional structure and location of generated concepts. Our results can be interpreted in terms of local minimization of between-category similarity: By distributing the generated category away from members of the experimenter-defined category, participants may increase the degree of between-category distance without drastically altering the

[4] Prior to plotting, data were also processed using a Gaussian filter with $\sigma = 0.8$.

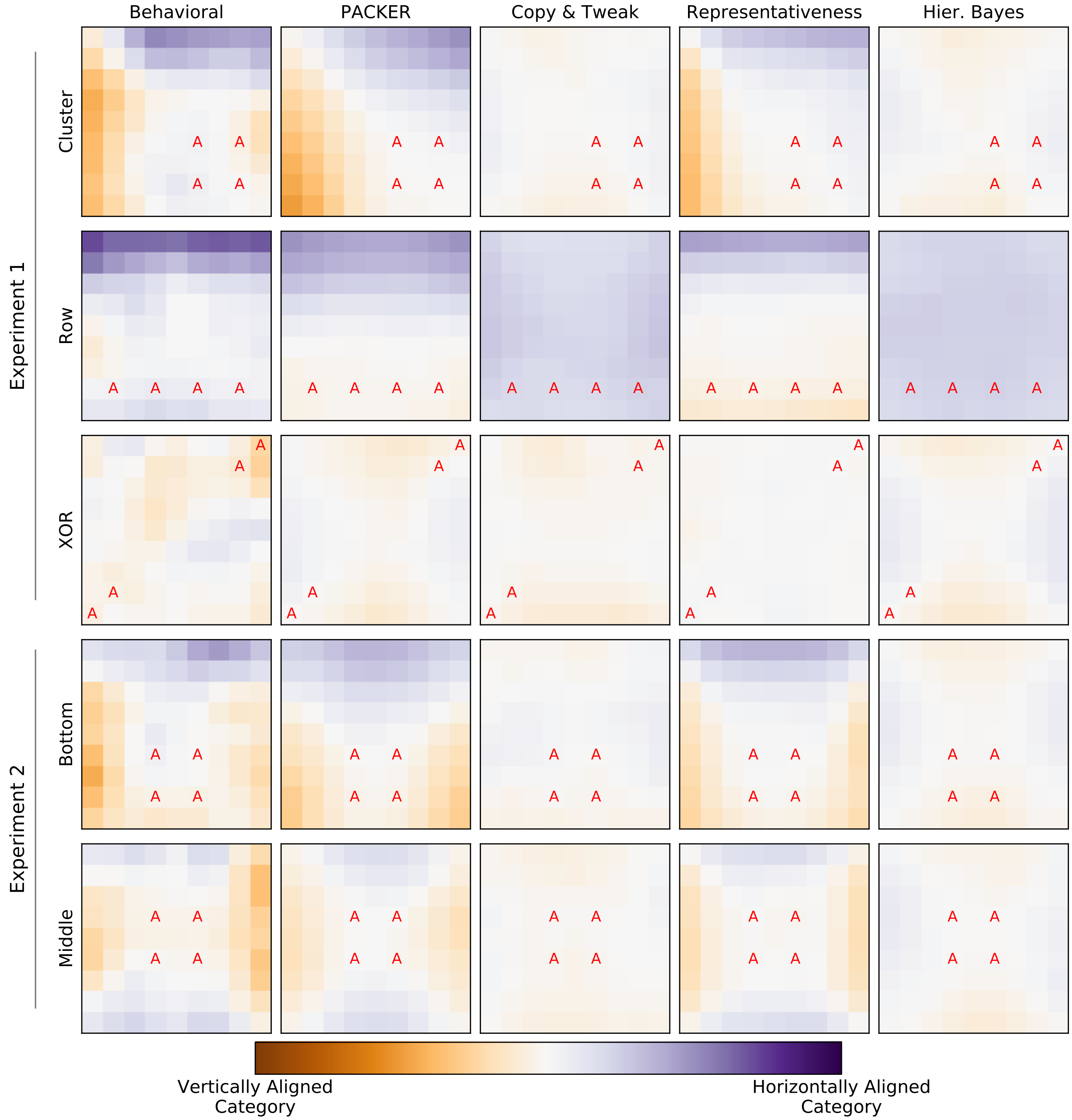

Figure 14: Behavioral and simulated range difference gradients. Each panel shows, for each stimulus, the dimensional orientation of the categories it was generated into: vertically aligned ‘columns’ (orange) versus horizontally aligned ‘rows’ (purple).

degree of within-category similarity.

To explore how well the PACKER, copy-and-tweak, representativeness, and hierarchical Bayesian models explain our findings, we conducted simulations using an individual-differences approach. As noted in Section 6.2, row- and column-like categories can be produced by each model through changes to the weighting of each dimension. Given this information, we may use the models to simulate each participant's generation separately, with the importance of each dimension set according to the relative range of the participant's generated category along each dimension.

In the PACKER and copy-and-tweak models, the attention weights, w, specify the importance of each dimension in the computation of similarity. While there exist methods to find the optimal attention weighting scheme given a classification (see Vanpaemel & Lee, 2012), for simplicity we assume that the 'Alpha' and 'Beta' categories are distinct along dimensions that the 'Beta' exemplars do not vary on. In this case, the weighting for a given participant can be computed as:

$$w_k = \frac{\exp\{-\theta_w \cdot \mathrm{range}(k)\}}{\sum_k \exp\{-\theta_w \cdot \mathrm{range}(k)\}} \tag{12}$$

where θ_w is a free parameter controlling how differences in range correspond to differences in weights (functioning similarly to the θ parameter in each of the models), and range(k) is the range of examples generated by the participant along dimension k. We used $\theta_w = 1.5$ in our simulations, though the results are robust and similar for other θ_w values. The resulting w values are thus inversely proportional to the range of generated categories along each dimension, with less range corresponding to greater weighting.

Unlike the PACKER and copy-and-tweak models, the representativeness and hierarchical Bayesian model's dimensional variances correspond to the assumed variance of generated categories along each dimension (rather than the inverse of the variance). Thus, a different transformation is appropriate for incorporating the weights computed in Equation 12. For the hierarchical Bayesian model, we computed the dimensional variances according to: $\lambda(1 - w_k)2$, where λ is a free parameter specifying the overall assumed variance of the domain, and 2 corresponds to the number of dimensions in our experiments[5]. Under this approach, evenly distributed weights correspond to an assumed variance of λ. Likewise, larger values of w, which are produced when the generated category is tightly distributed along one dimension, correspond to smaller assumed variances.

Each model was used to simulate each participant's generation independently, with the importance of each dimension set according to the participant's generated category. The other free parameters within each model were set as in Table 2. Every participant's generation was simulated 2,000 times; given the 305 participants tested across the two experiments, each model generated 610,000 categories in total. For comparison with our behavioral results, we then computed the range difference gradient identically as with the behavioral data. The results are shown in Figure 14.

As in the more traditional model evaluation analysis described above, the contrast models (i.e., PACKER and the representativeness model) provided a much closer match to our behavioral results than the copy-and-tweak and hierarchical Bayesian models. In all conditions, the contrast models distribute categories similarly to the behavioral data: Horizontally-aligned categories tend to be placed above and below members of the experimenter-defined category, and vertically-aligned categories tend to be placed to the sides. Conversely, because the copy-and-tweak and hierarchical Bayesian models are insensitive to category contrast, these models do not produce any systematic patterns of association between category location and distributional structure. The sole exception is within the 'Row' condition of Experiment 1, in which the majority of participants generated a 'Row'-like category, widely distributed along the X-axis but not the Y-axis. In these cases, both models are initialized with weights that produce Row categories, but because category contrast is not considered, categories are uniformly generated across the entire domain, rather than concentrated within the upper-regions as observed behaviorally.

[5] This calculation applies only in two-dimensional domains, where $w_2 = 1 - w_1$.

7 Experiment 3: Contrasting Contrast

Experiments 1 and 2 clearly established the importance of contrast in category generation. Follow-up model-based analyses illustrated that both contrast models account for participant performance better than models that do not take into account contrast. We also found more support for the hypothesis of contrast as representativeness versus contrast as exemplar dissimilarity across the different Alpha conditions.

While the representativeness model fits better than PACKER, and in spite of their fundamentally different structures, both contrast models make qualitatively similar predictions for Experiments 1 and 2. To illustrate this, Figure 15 shows how the models' generating probabilities for the first Beta exemplar are similar across both contrast models.[6] Across the five Alpha conditions of Experiments 1 and 2, the first Beta exemplars according to the contrast models were generally more probable near the corners and edges. This is particularly the case for the representativeness model, which almost exclusively predicts generation at the corners, while PACKER can predict generation in the center of the feature space in the XOR condition.

The effect of contrast for both models is similar in that they both tend to produce exemplars that are distant from the Alpha category. However, this is realized in different ways for the two models. The representativeness model does not explicitly use a measure sensitive to distance from the location of exemplars in the contrasting category. Instead, it is the assumption of unimodal distributions that results in the model producing novel exemplars dissimilar to learned exemplars. Consequently, it is possible to design a configuration of Alpha exemplars where the representativeness model predicts novel exemplars should be *similar* to the learned exemplars. However, PACKER cannot predict this because it assumes that the similarity weight for contrasting exemplars is negative.

In this section, we focus on a different category type – the Corner category – that contrasts the two models of contrast. Here, each Alpha exemplar is located at the corners

[6] Model parameters are set to the optimized values from Table 2.

of the feature space (see Figure 16a). We first identify the predictions of each model in this condition. Second, we test these contrasting predictions in a behavioral experiment. We then conduct a model-based analysis of the two contrast models.

In the Corner condition, the representativeness model estimates the underlying Alpha category distribution to be centered on the feature space, with probability densities highest in the center and lowest in the corners. The representativeness of novel Beta exemplars generally decreases with the likelihood of the underlying Alpha category distribution. Despite the Alpha exemplars being in the corners, the predicted mean will be near the center of the Alpha category (the center of the space; See 16b). As such, the Alpha likelihood is lowest in the corners, and thus, the representativeness of the Beta category will be largest in the corners.

Conversely, PACKER predicts a completely different result. With PACKER, the probability of generating a Beta exemplar increases with the dissimilarity from all Alpha exemplars. This results in Betas generated in the center of feature space, where the candidate exemplars are as far as possible from the Alpha exemplars (Figure 16c).

Another aspect of the contrast in the models' predictions is the role of the boundaries of feature space. The representativeness model prefers to generate categories at the boundaries of feature space. PACKER is less sensitive to boundaries. Boundaries tend to be where distance to the contrasting category exemplars is maximized. But, as in cases like the Corner condition (Figure 16a), this is not always the case; Beta exemplars should be as far from the Alphas as possible, which is maximized in the center of feature space.

Thus, in this experiment, we also examine how making one feature toroidal (e.g., line orientation) affects the prediction of each model. The representativeness model prefers to generate categories on the feature that still has a boundary. PACKER, in contrast, should continue to generate categories that are most distant from the learn category regardless of the presence or absence of feature boundaries – in this condition it would be the center of the bounded feature. Experiment 3 manipulates whether one feature is

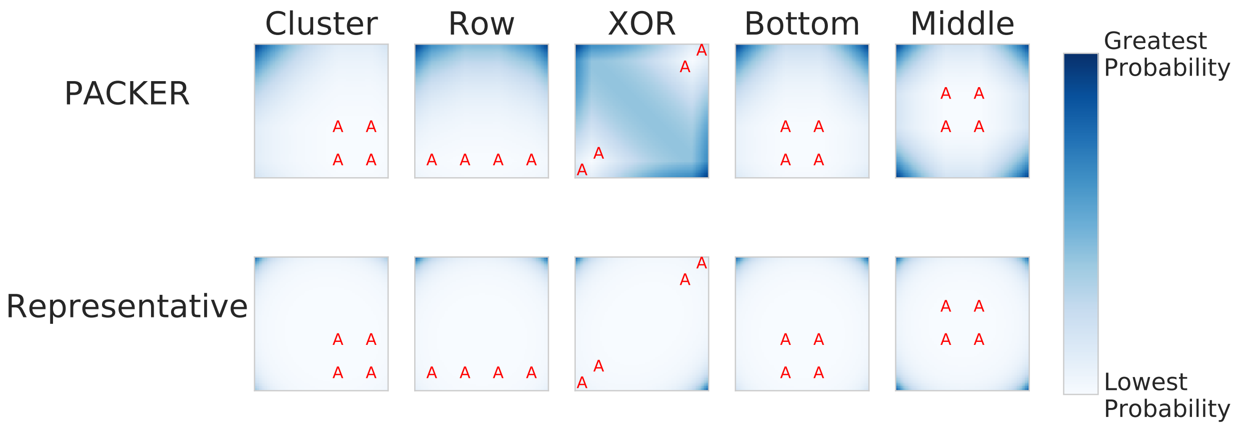

Figure 15: Contrast model predictions indicating probability of generating the first Beta exemplar for each Alpha condition. Darker shades of blue indicate higher probabilities (normalized within each plot).

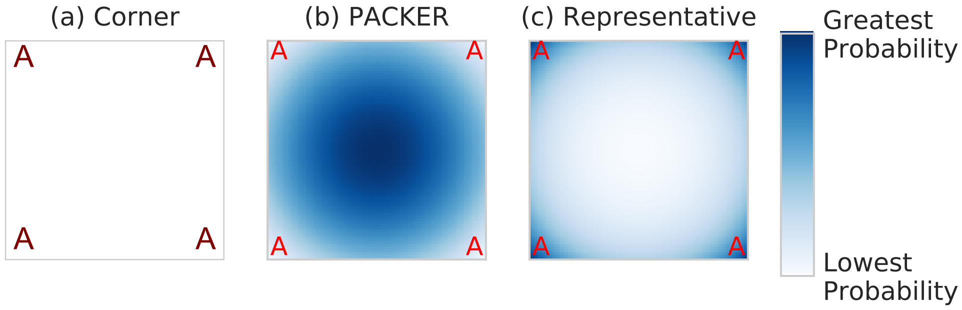

Figure 16: Alpha exemplars for the Corner condition (a) and the predictions from contrast models PACKER (b) and the representativeness model (c) with square stimuli.

bounded or toroidal (always displayed as the Y-axis). One condition uses the same stimuli as before, squares varying in lightness and size, both of which are bounded features. The other condition use a new set of stimuli: circles that vary in size (a bounded feature) and orientation (an unbounded feature; Figure 17a). Before describing the details of the experiment, we explain in detail each model's predictions given the Corner Alphas for a feature space where one feature is bounded and the other is toroidal.

Compared to stimuli with two bounded features, PACKER makes a similar prediction in generating the first Beta exemplar when one feature is unbounded, with the curious exception that the highest density regions appear compressed along the toroidal feature. Due to the toroidal nature of orientation, the maximum distance from the Alpha exemplars is half of the full range of the feature – for an Alpha exemplar oriented at 0 radians, the furthest distance to another exemplar is π radians.

For the representativeness model, adding a toroidal feature changes the probability distribution for the underlying Bayesian model. For the toroidal feature, space wraps back to 0 radians once it reaches 2π (we have conducted a change of variables where $y \to y \bmod 2\pi$, where mod is the modulus operator). As we cross 2π and reach 0 again, the remaining probability density also gets wrapped around, meaning that the density at 2π is added to 0 (and $2\pi + \epsilon$ to ϵ, and so on). This is called the *wrapped* Gaussian distribution (Mardia & Jupp, 1972). Note that the more familiar distribution for circular features, the Von Mises distribution, is an approximation to the wrapped Gaussian distribution. See Appendix B for implementation details of the representativeness model in a bounded and toroidal feature space.

The consequence of employing a wrapped Gaussian is presented in Figure 17c. Edge effects along the unbounded feature (displayed as the vertical axis) are no longer present since the edges themselves no longer exist. Instead, the probability of generating the first Beta exemplar is bimodally distributed at the edges of the bounded feature (displayed as the horizontal axis). Thus, the representativeness model and PACKER make orthogonal predictions for how a toroidal dimension should affect category generation.

In this experiment, we test human category generation in two cases where PACKER and the representativeness model contrast. The representativeness model predicts categories should be distributed in the boundaries of the bounded features. PACKER predicts category exemplars should be centered in space and decay exponentially over bounded dimension(s).

7.1 Participants, Materials, and Procedure

We recruited 89 participants from Amazon Mechanical Turk who were randomly assigned to the square stimuli ($N = 46$), or to the circle stimuli ($N = 43$). The square stimuli were exactly the same as in Experiments 1 and 2, but with Alpha exemplars in the corners of the feature space. The circle stimuli were made up of unfilled circles with a radial line extending from its center to its edge. These stimuli varied along two features: their diameter within the range of [3.0, 5.8] cm (inclusive), and their orientation in the range $[2.587, 2\pi + 2.587]$ radians (inclusive), with 0 radians indicating a completely horizontal line and orientation increasing in an anti-clockwise direction. The line widths of the circles and radial line were set to 0.05 cm. Similar to Experiments 1 and 2, participants first observed a training phase where four Alpha exemplars were presented across three blocks followed by the generation phase, and the features values are divided into 9 discrete steps. The training condition was the Corner category for all participants.

7.2 Results and Discussion

Unlike previous experiments, Experiment 3 allows us to distinguish between PACKER and the representativeness model beyond their quantitative fits by identifying specific qualitatively diagnostic markers. Specifically, if PACKER is a better account of the data, we would expect a relative decrease in the range of the entire Beta category along the unbounded feature compared to the bounded feature. In addition, we would expect to see a preference for first-generated Beta exemplars away from the boundaries of the feature space – that is, the probability of generating an exemplar would be negatively correlated with its distance from the center for bounded features. These predictions are correspondingly reversed if the representativeness model offers a better account of the data.

Figure 18 presents the distributional statistics for both square and circle stimuli. Participants produced circle stimuli with significantly lower Y-range (lower for the unbounded feature) as compared to the square stimuli ($t(87) = 4.94$, $p = .001$), which supports PACKER. There are no significant differences between the distributional statistics of each generated category between square and circle stimuli (Figure 18).

The distributions of generated first-Betas (Figure 19) indicate further qualitative support for PACKER. In particular, the probability of generating the first Beta exemplar for circles along the bounded dimension is significantly negatively correlated with distance from the center ($r(3) = -.95, p = .01$). Although the correlations for squares ($r(3) = -.50, p = .40$ and $r(3) = -.31, p = .61$ for X and Y dimensions respectively) were not significant, they were in the directions predicted by PACKER. Mean probabilities of generating first-Betas are presented in Figure 20. This analysis was excluded for the toroidal dimension of the circle stimuli because the boundaries (and consequently a center) are not defined for this feature.

Model generation probabilities of the first Beta exemplars do not qualitatively differ from earlier predictions even after fitting to the current dataset. Figures 21a-b and 21c-d show the model predictions for square and circle stimuli respectively. While the predictions for the representativeness model for circles appears different from earlier predictions (i.e., in Figure 17c), note that our key expectation for this model was still met – specifically, it predicts the highest first-Beta generation probabilities at the edges of the bounded features.

The contrast models were fit by optimizing them to only Experiment 3. Fit statistics and parameter values are presented in Table 3. Although not our main focus in this section, we also provide the statistics for the classical models – unsurprisingly, they do

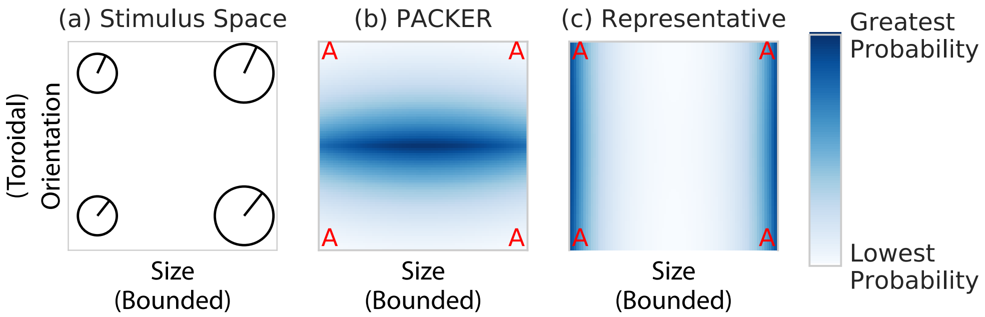

Figure 17: Circle stimulus domain (a) and predictions from contrast models PACKER (b) and the representativeness model (c). The unbounded feature (orientation) is represented on the Y-axis, while the bounded feature (size) is represented on a X-axis. As with the square stimuli, feature and direction assignment were counterbalanced across participants (in the displayed example, circles are rotated clockwise up the Y-axis).

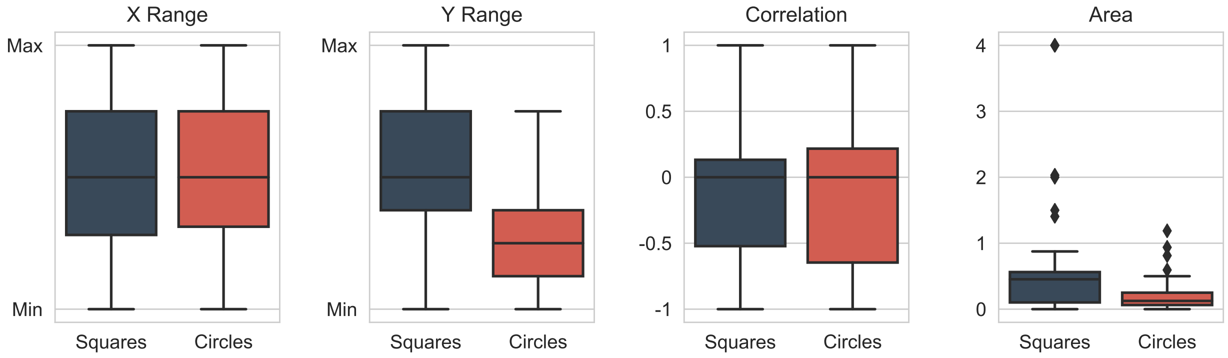

Figure 18: Box-plots of the distributional statistics from the categories generated in Experiment 3. Boxes depict the median and quartiles of each condition, with whiskers placed at 1.5 IQR. All points outside this region are marked individually. For circle stimuli, the Y Range represents orientation and the X Range represents size.

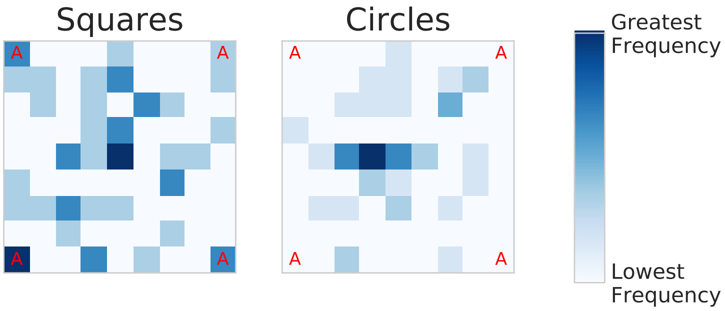

Figure 19: Heatmaps indicating the relative frequencies of generating the first Beta exemplar in the Corner condition.

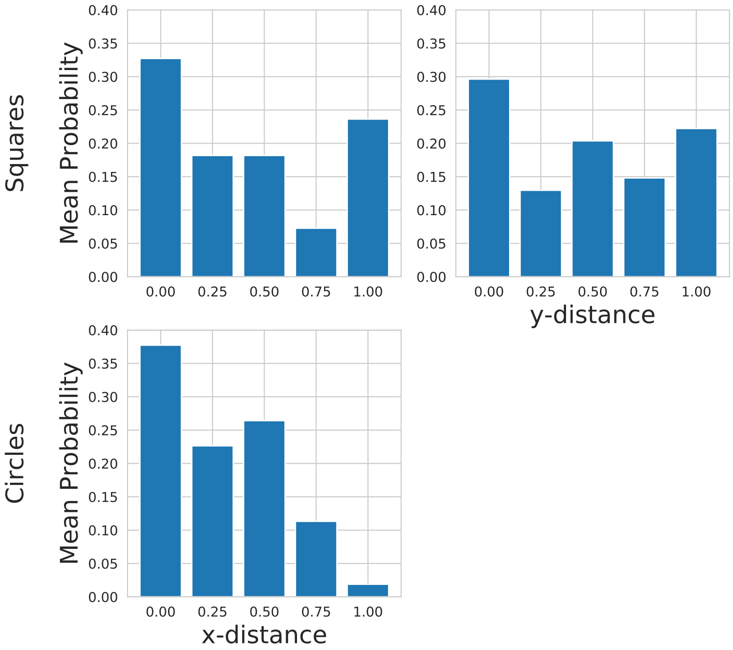

Figure 20: Empirical mean probabilities of generating first Beta exemplars at each point of distance from the center of the feature space. Distances are normalized such that 0 represents the center and 1 represents the edge. Plot for the toroidal feature of circle stimuli is not presented because a center is not defined for this feature.

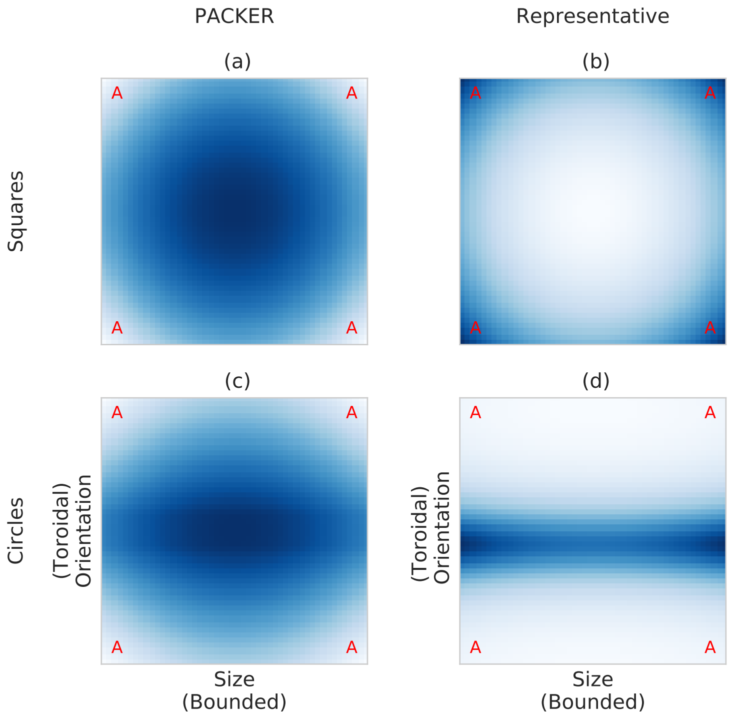

Figure 21: Heatmaps indicating the probabilities of each contrast model generating the first Beta exemplar in the Corner condition for both square and circle stimuli. Models have been fit to data from Experiment 3.

Table 3: Results of model-fitting to the combined datasets from Experiments 1, 2 and 3. Smaller AIC values correspond to better model fits (adjusted for number of parameters).

PACKER	**Copy & Tweak**	**Representativeness**	**Hierarchical Bayesian**
$AIC = 2954$	$AIC = 2963$	$AIC = 2971$	$AIC = 3025$
$L = -1474$	$L = -1479$	$L = -1481$	$L = -1508$
$c = 2.08$	$c = 3.73$	$\kappa = 0.00$	$\kappa = 0.29$
$\theta_c = 0.87$	$\theta_c = 0$ (fixed)	$\nu = 4.07 \times 10^{20}$	$\nu = 1.00$
$\theta_t = 2.18$	$\theta_t = 2.57$	$\lambda = 0.14$	$\lambda = 0.00$
		$\theta = 0.73$	$\theta = 7.50$

Table 4: Frequencies of individual participants best fit by each contrast model broken down by stimulus type.

Stimulus Type	**PACKER**	**Representativeness**	Total	Two-tailed Binomial test p-value
Squares	28	18	46	$p \approx 0.18$
Circles	35	8	43	$p < 0.001$
Total	60	26	89	$p < 0.001$

not perform as well as the contrast models. In line with our qualitative observations, PACKER emerges as the best-fitting model in aggregate.

Table 4 presents frequencies of the best-fitting model for participants in each condition. Overall the representativeness model was significantly worse at capturing the results of Experiment 3 (it fit 26 out of the 89 participants better than PACKER; two-tailed Binomial test $p < 0.001$). Further, this is mostly due to the representativeness model capturing participants in the unbounded circle condition worse than PACKER (8 of 43 better fit by representativeness; two-tailed Binomial test, $p < 0.001$). The representativeness model fit the bounded Square condition better, but still worse than PACKER (18 out of 46; two-tailed Binomial test $p = 0.18$).

Ultimately, these results suggest that PACKER's exemplar dissimilarity account of contrast, as opposed to the representativeness account, offers a better explanation of category generation for the Corner category type. On almost every measure of analysis in this section, PACKER outperformed the representativeness model, indicating that the better fit of the latter model demonstrated in Experiments 1 and 2 is not general to all conditions.

8 General Discussion

The sensory impression of every stimulus and event is unique. Grouping distinct patterns of sensory information into categories is a fundamental task solved by the mind. Most work has focused on how people learn new categories that are provided to them through unlabeled, partially labeled, or fully labeled examples. How were these categories first determined? Some natural categories are likely to be the result of regularities in the dynamics of our environment. But, these are only a subset of the categories that people learn. Other categories, such as tools and ideas, were generated by people over time. What basic principles underlie how people generate categories?

While the bulk of prior research on categorization has focused on the classic finding that generated concepts tend to be distributionally similar to known concepts, there has been little work addressing the role of contrast in category generation: How is it that people are able to create something *different* from what is already known? We developed two novel models, each incorporating a different conceptualization of category contrast. One model is PACKER, which is an exemplar-based model that formally specifies the role of contrast in generation as exemplar dissimilarity. Specifically, the model proposes that categories are represented as exemplars in a multidimensional psychological space, and generation is constrained both by within-category and between-category similarity: Exemplars belonging to the same category should be similar to one another, and exemplars belonging to different categories should not be similar to one another. The second model is a novel hierarchical Bayesian model with a representativeness mechanism. This model generates exemplars that are more representative of (i.e., has greater relative evidence from) the novel category compared to the learned category.

We reported two experiments demonstrating systematic effects of category contrast in category generation. Members of participant-generated categories tended to be highly dissimilar from members of previously-learned categories, and were usually more similar to one another than to members of other categories. We also observed broad interdependence between the distributional structure (feature variance, correlation) and physical instantiation (location within the feature space) of generated categories: In Experiment 2, we found that the unoccupied regions of the domain influenced the distributional structure of categories, and in both experiments we observed that participants distributed their generated categories to increase contrast with what was already known.

We conducted simulations comparing the contrast models' account of our results to the classical proposals for category generation: a "copy-and-tweak" model (realized as a variant of PACKER with no sensitivity to category contrast), and a hierarchical Bayesian model designed to explain the classic distributional similarity effect. In all simulations, we found that the contrast models captured a previously unexplained and unexplored aspect of human category generation. In particular, by measuring PACKER's fit as a function of its prioritization of within- and between-category similarity, we observed that considering either constraint exclusively results in a relatively low-quality account. Instead, PACKER's best results were obtained when both constraints are considered, indicating that human learners do not generate novel concepts exclusively on the basis of within-category similarity or between class-contrast. This finding mirrors our behavioral results and demonstrates that both constraints influence generation when explained with an exemplar model.

While we found that the representativeness model was found to be a better fit to data from Experiments 1 and 2, both contrast models essentially made qualitatively similar predictions across all conditions in those experiments. In Experiment 3, we collected behavioral data on a category generation condition that was designed to qualitatively distinguish the two contrast models. Specifically, in this condition where exemplars were located at the corners of the feature space, PACKER predicted the initial generation of centrally-located novel exemplars while the representativeness model predicted novel exemplars similarly located at the corners. Interestingly, data from this experiment were better fit, qualitatively and quantitatively, by PACKER than the representativeness model.

8.1 Similarity and Contrast in Cognition

We propose category contrast as a primary constraint in categorization. For categories to be useful, they should not all be identical, or in other words, they should be different. Thus, a newly generated category should be different than pre-existing categories. Beyond its role in category generation specifically, category contrast is also of fundamental importance in categorization more broadly. All other factors held constant, new categories are easier to learn if they are dissimilar to members of other categories, and knowledge of highly distinct categories is applied more accurately than that of ill-defined categories (Ashby, Boynton, & Lee, 1994; Imai & Garner, 1965). Likewise, basic-level categories (Rosch et al., 1976) are thought to be abstracted in order to maximize within-category similarity while minimizing between-category similarity. Finally, the act of forming category representations affects similarity judgments about category members and nonmembers, with category members being viewed as more similar to one another than members of other categories (Goldstone, 1994, 1996; Goldstone, Lippa, & Shiffrin, 2001).

Beyond the traditional categorization literature, one can find instances of the trade-off between within and between-class similarity in linguistic categories over perceptual dimensions. For example, Regier et al. (2007) showed that the partitioning of color categories reflects such a trade-off in a psychological space – colors are partitioned into groups with members that are viewed as highly similar to one another yet distinct from other colors. A similar trade-off can be observed in phoneme categories. Different exemplars of the same phoneme must be similar to one another, while contrasting from other phonemes, such that a listener can infer the appropriate phoneme. This pattern has been found and modeled in the natural acoustics of American English vowels (Feldman, Griffiths, Goldwater, & Morgan, 2013; Hillenbrand, Getty, Clark, & Wheeler, 1995). As linguistic categories must have been created at some point in human history, it is revealing that the constraints of emulating distributional structure across categories and having categories contrast from one another still bias human category generation today.

The dual forces of within-class similarity and between-class contrast influence cognitive functions in a wide variety of domains. The PACKER model is notable in that it explicitly interprets this trade-off within the domain of categorization, and allows us to begin to understand the relatively understudied processes involved in category generation through our more well-developed tools for understanding human categorization.

8.2 Implications for Creative Cognition

Although the focus of this article has been to address the role of contrast in category generation, our findings and approach have relevant implications for research in creative cognition. A central focus of the creative cognition approach has been to explain acts of creativity in terms of the mental representations and processes that are commonly studied in cognitive psychology and cognitive science (Finke, Ward, & Smith, 1992; S. M. Smith et al., 1995). However, unlike other fields in the study of cognition, creative cognition research rarely employs quantitative models to evaluate the explanatory value of such representations and processes. Our modeling results provide a concrete example of how formal approaches may be used to gain insight into the nature of creative cognition.

In addition to demonstrating the utility of formal modeling for studying creative cognition, the contrast models here specifically offer an additional interpretation of some of the field's most central findings. For example, perhaps the most foundational principle from this literature concerns the limiting influence of prior knowledge: Individuals create new categories composed of features from existing classes, and what is created can be influenced drastically through the introduction of cues or examples (Marsh et al., 1999;

S. M. Smith et al., 1993). In this paper, we have identified another important aspect of the constraining influence of prior knowledge: What is generated cannot be the same as what is already known. Further, there is systematicity in how generated categories differ from prior knowledge. The results of our simulations suggest that this conceptualization of difference can be addressed in at least two different ways. Specifically, simulations with PACKER demonstrate that the constraining influence of difference is concisely explained in terms of a trade-off between within-category similarity and between-category dissimilarity. Conversely, the representativeness model shows that this influence can also be the result of enhancing the representativeness of generated exemplars to their category.

PACKER may offer an additional interpretation of existing accounts of creative generation. Most notably, a leading account within the creative cognition literature, the Path of Least Resistance (Ward, 1994, 1995), also explains generation in terms of an exemplar-based retrieval process. This account was designed to explain the creative generation of natural categories (e.g., new species of plants and animals) and as a result relies strongly on the hierarchical organization of these categories: Individuals are thought to retrieve an example of the higher-level category being generated (e.g., *bird* may be retrieved from the category *animal*), and then systematically alter what was retrieved to make something new. As the PACKER model does not assume knowledge is hierarchically organized (this is true of the exemplar view more broadly, see G. L. Murphy, 2016), the model may be viewed as a formal instantiation of the Path of Least Resistance for application in a traditional artificial categorization domain (when there is no established hierarchy of categories). PACKER's success in explaining generation within an artificial domain motivates future work exploring the nature of category contrast within a more naturalistic setting.

The broader study of creativity currently involves a wide breadth of different approaches (for a review, see Kozbelt, Beghetto, & Runco, 2010), such as those based on free association (Mednick, 1962) and conceptual combination (Estes & Ward, 2002;

G. L. Murphy, 1988). In addition, recent work in the machine learning literature has explored using neural networks to address the overall problem of creative generation (e.g., Chen et al., 2016; Goodfellow et al., 2014; Ho & Ermon, 2016; Kingma et al., 2016). In contrast to these varied investigations, we reduced our focus by studying a highly complex behavior (category generation) as it applies within a well-established domain (artificial category learning). We hope that future work incorporates and highlights the importance of contrast into theories of creativity and adapts the discussed contrast models to work in state-of-the-art creative generation methods.

8.3 Implications For Categorization

Categorization research addresses the representations and processes that underlie the learning and use of categories. Category learning tasks are generally about figuring out which items belong to which category. Once learned, categories are generally used to classify new stimuli and to make inferences beyond the available information. Our work is fairly unique in that people learn a category through positive examples and then create another category that would make sense in the domain. Such scenarios have direct application to real-world situations. For instance, consider a rock musician in the 1970s who wants to write new types of rock music. Given the diversity existing rock music that the musician is aware of, the musician must identify features of this new sub-genre that are unique to that sub-genre (e.g., introduce instruments that may not typical to rock music at the time, such as accordions or synthesizers). In the present studies, we have learned something about the form that such expectations on new categories are likely to take.

We can think of the category generation task in our studies as asking a person to formulate an idea about what set of items in the domain are most interestingly *not* members of the original category. To meet this condition, the items must take some form of coherence that aligns with that of the original category and some form of distinctiveness relative to the original category. Reflecting the basic level of organization in natural categories, it makes sense to generate a set of items that possess strong within-category coherence (by importing or systematically transforming the internal structure of the original category) as well as strong between-category differentiation (by creating maximum contrast with the original category, be it through exemplar-based dissimilarity or the maximization of exemplar-category representativeness). In this sense one can see the patterns of performance in the category generation task as recapitulating the order of semantic organization.

The current work also suggests exciting directions for related investigations into unsupervised categorization. While the unsupervised categorization literature is primarily interested in the generation of categories within a set of observed exemplars where no prior categories are learned (e.g., Pothos et al., 2011), our category generation studies are focused on the production of novel category exemplars themselves. Despite this distinction, both types of categorization research are ultimately interested in the question of how categories can be formed. The results of our current work indicate that the formation of categories is constrained by some measure of contrast. To our knowledge, this explicit investigation of category contrast has not yet been attempted in the unsupervised categorization literature.

Interestingly, the unidimensional sort bias that is commonly observed in unsupervised categorization (Ahn & Medin, 1992; Imai & Garner, 1965; Milton & Wills, 2004), where participants sort unlabeled family resemblance category exemplars by focusing on a single feature, was not consistently observed in our tasks. In our experiments we only observed this bias in the row condition of Experiment 1, whereas in other conditions participants generated categories that typically varied along both dimensions. This may be unsurprising at face value, since participants in conditions other than the row condition were trained on a prior category that varied on both dimensions. However, these results place constraints on how broadly applicable the unidimensional bias is to different types of category construction. Additionally, the unidimensional sort bias is observed in category construction tasks in which participants develop a category-level organization of a provided domain while category generation tasks involve generating a new category relative to one that was just acquired (one might say that participants are forging a domain from a category rather than categories from a domain). Since the more creative task of generating a category from scratch (with respect to an established category) does not invoke the unidimensional bias, this suggests that a more creatively demanding version of a sort task might also reduce the bias.

8.4 Exemplar Dissimilarity and Representativeness in Categorization

Although both PACKER and the representativeness model perform better than the classical models, they appear to be capturing fundamentally different aspects of contrast in categorization. Specifically, the representativeness model excelled in predicting categories generated by participants in Experiments 1 and 2. Conversely, PACKER was the best-fitting model in accounting for categories generated from the Corner category type, especially when the stimuli had a toroidal feature.

While the contrast models were born quite naturally out of the corresponding classical models, it is worth considering if similar benefits in prediction would be seen if the contrast mechanisms were applied to different classical models. Although there is no clear way to adapt the exemplar dissimilarity mechanism to the current hierarchical Bayesian model, it is a fairly straightforward process applying the representativeness mechanism to an exemplar model. Specifically, by treating the similarity measure (Equation 1) in the current implementation of copy-and-tweak as a density estimate for the representativeness mechanism (Equation 4), we gain a new model of representativeness that is based on exemplar similarity instead of a multivariate Gaussian likelihood.[7]

How well does a representativeness model that uses exemplar similarity to represent

[7] An analogous implementation of the representativeness mechanism to PACKER results in both θ_c and θ_t adding to constant value that is independent of the similarity to contrast and target exemplars respectively. Consequently, it is formally equivalent to a copy-and-tweak model with representativeness.

the category's density over features explain our experimental results? We find that for data from experiments 1 and 2, this model substantially outperforms its classical counterpart, but not to the extent of PACKER or the representativeness model. Specifically, the model fits here yielded $L = -4815$ and $AIC = 9633$ for the entire set of data from Experiments 1 and 2 and $L = -3474$, $AIC = 6953$ for the same data excluding the first trials (cf. Table 2). When we fit the model to data from Experiment 3, we find that $L = -1486$, $AIC = 2975$, which indicates a poorer fit than its classical counterpart (cf. Table 3). Overall, the representativeness model performed best when generating categories where previous category examples were not at the edges of the feature space. Conversely, PACKER performed better when the previous category was at the extremes of the features or if a feature is bounded. These results suggest that the predictive advantages to including the representativeness mechanism is not limited to a hierarchical Bayesian model, but can also be found when applied to an exemplar model. Although its performance is not as strong as either of the fully-developed contrast models, there can be a benefit to the representativeness mechanism that is independent from any interaction with a hierarchical Bayesian framework. Ultimately, it appears that there is no single model that can capture the entirety of our empirical data.

8.5 Limitations and Future Directions

Although successful in explaining our results in all three studies (albeit not as well as the representativeness model for Experiments 1 and 2), PACKER does not provide a full account of what is known about category generation. Most notably, in this paper we have not evaluated the model's ability to explain the classic finding that generated categories tend to share distributional commonalities with previously learned categories (see Jern & Kemp, 2013; Ward, 1994). While we successfully replicated this effect in Experiment 1, we also found that its influence was limited in comparison to the fundamental constraints imposed by category contrast. Even within Experiment 1, we found systematic inconsistencies: by generating exemplars into unoccupied regions of the space, participants who learned an 'XOR' category, composed of members that are widely distributed along both features and are positively correlated in space, tended to generate categories with an opposite (negative) correlation. More generally, PACKER inherits the strengths and weaknesses of exemplar models of categorization: It provides a simple and flexible model that explains many results, but deviates systematically from human performance in some cases.

Nonetheless, these classic effects are a core element of the phenomenology of category generation, and PACKER does not include any mechanisms that explain them. Instead, through the development and evaluation of the PACKER model, we have sought to add new elements into such a phenomenology: The broad and strong influence of category contrast, and the interdependence between category location and distributional structure. It may be possible to combine the hierarchical Bayesian approach proposed by Jern and Kemp (2013) with PACKER's underlying claims to obtain a "best of both worlds" model, capable of explaining the role of contrast in category generation, as well as the emulation of distributional structure. However, as noted in the introduction, the incorporation of category contrast is antithetical to the core principles of a traditional, semi-conjugate Bayesian approach. This suggests that category generation is a fundamentally different computational-level problem (different from those posed by Jern & Kemp, 2013; Kemp & Jern, 2014), which is supported by the representativeness form of their model performing much better.

Characterizing that problem and conducting a rational analysis is an important direction for future research. To that aim, we plan to explore the connection between exemplar modeling as an Importance Sampling approximation (Shi, Griffiths, Feldman, & Sanborn, 2010), and see what sort of computational-level problem PACKER approximates. Once formalized in probabilistic terms, it should also be straightforward to incorporate distributional factors into the model. This would unite PACKER with the Bayesian representativeness model, and they would differ in terms of their assumptions about how people represent category distributions (as exemplars or prototypes, respectively). Alternatively, it may be possible to integrate the core principles of either model of contrast into other categorization models (e.g., Kurtz, 2007; Love et al., 2004; D. J. Smith & Minda, 2000).

The present work focuses on the influence of contrast on categorization, primarily by exploring two different conceptualizations of contrast. However, throughout this paper we have only explored one type of category generation problem, that is the generation of new categories in an artificial domain. It is possible that the strength of the influence of contrast on category generation decisions varies depending on the nature of the category generation problem, for instance, in situations where the boundaries of the domain are not clearly defined. To provide some intuition for this, consider the following example: an entomologist is asked to draw a new type of insect that they have never seen. Clearly, features of the new insect will be constrained by the definition of an insect: it will be an arthropod with six legs and a relatively hard exoskeleton. However, there are no strict limitations on other features such as how large the insect needs to be, or if the insect should have a particular type of pattern on its body. In this scenario, it is plausible that the entomologist will draw an insect that is fairly similar to insects that currently exist (e.g., small in size with a camouflage pattern that matches its environment) as opposed to something completely different (e.g., a horse-sized beetle with roman numerals on its back), indicating that contrast plays only a minor role in the decision making process. In this situation it appears that there is a combination of clearly defined constraints on the domain boundary that are imposed by the external environment (insects must be arthropods with six legs) but also ill-defined constraints that are imposed by the observer (insects are probably small and have patterns on their body that suit the environment). However, the relation of these constraints (and the interactions between them) to the role of contrast is not currently well understood.

The definition of the domain boundary is only one possible factor that could influence the effect of contrast on categorization. It is also plausible that other conditions such as the nature of the instruction ("Generate a new category that is Not Alpha" vs "Generate a new category that is Beta") can affect the extent to which contrast influences categorization. Though our current contrast models can allow the influence of contrast to vary, they do not make any predictions regarding the specific conditions under which contrast can vary. Consequently, addressing this issue would be an immensely promising avenue for future research.

9 Conclusions

The generation of new concepts and ideas is a highly interesting topic, but it is difficult to study in a controlled experimental environment. In this paper, we have provided such an examination of category generation as it applies within an artificial categorization experiment. Extending the literature on creative cognition, our experiments provide a detailed picture of the role of category contrast in generation: People seek to create concepts that are distinct from what they already know, and the nature of what is created can be influenced by what does not yet exist. Our simulations with traditional exemplar models, as well as a hierarchical Bayesian model, provide strong support for the claim that category contrast is of fundamental importance to categorization.

10 Acknowledgments

Previous versions of this work were presented at the Thirty-Ninth Annual Conference of the Cognitive Science Society and Forty-Ninth Annual Meeting of the Society for Mathematical Psychology. Support for this research was provided by the Office of the VCRGE at the UW-Madison with funding from the WARF and the Vilas Life Cycle Fellowship. We thank Charles Kemp for feedback on a previous version of this manuscript. We also thank him and Alan Jern for providing code and data.

References

Abbott, J. T., Griffiths, T. L., & Reiger, T. (2016). Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences*, *113*(40), 11178–11183.

Abbott, J. T., Heller, K. A., Ghahramani, Z., & Griffiths, T. L. (2011). Testing a Bayesian model of representativeness using a large image database. In *Advances in neural information processing systems* (pp. 2321–2329).

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81–121.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, *55*(1), 11–27.

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, *82*(5), 910-944.

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, *63*(4), 516–556.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, *11*(4), 629-654.

Berger, J. (2016). *Contagious: Why things catch on.* New York: NY: Simon and Schuster.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Elbaum.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett

(Eds.), *Advances in neural information processing systems 29* (pp. 2172–2180). Curran Associates, Inc.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, *30*(3), 353–362.

Conaway, N. B., & Kurtz, K. J. (2016a). Generalization of within-category feature correlations. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2375–2380). Austin, TX: Cognitive Science Society.

Conaway, N. B., & Kurtz, K. J. (2016b). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic Bulletin & Review*, 1–12. doi: 10.3758/s13423-016-1208-1

Estes, Z., & Ward, T. B. (2002). The emergence of novel attributes in concept modification. *Creativity Research Journal*, *14*(2), 149–156.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications.* Cambridge, MA: MIT press.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608-628.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes,

N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2672–2680). Curran Associates, Inc.

Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive psychology*, *103*, 85–109.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind.* Oxford, England: Oxford University Press.

Hahn, U., & Warren, P. A. (2009). Perception of randomness: why three heads are better than four. *Psychological Review*, *116*, 454–461.

Hidaka, S., & Smith, L. B. (2011). Packing: a geometric analysis of feature selection and category formation. *Cognitive Systems Research*, *12*(1), 1–18.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 4565–4573). Curran Associates, Inc.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 297–307.

Imai, S., & Garner, W. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, *69*(6), 596.

Jern, A., & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, *66*(1), 85–125.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. In *The concept of probability in psychological experiments* (pp. 25–48). Springer.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Kemp, C., & Jern, A. (2014). A taxonomy of inductive problems. *Psychonomic bulletin & review*, *21*(1), 23–46.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 4743–4751). Curran Associates, Inc.

Kozbelt, A., Beghetto, R. A., & Runco, M. A. (2010). Theories of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The cambridge handbook of creativity* (p. 21-48). New York: Cambridge University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552-572.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, *19*(2), 189-223.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, *111*(2), 309–332.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of verbal learning and verbal behavior*, *23*(2), 250–269.

Mardia, K. V., & Jupp, P. E. (1972). *Statistics of directional data*. London: Academic Press Inc.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613.

Marsh, R. L., Ward, T. B., & Landau, J. D. (1999). The inadvertent use of prior knowledge in a generative cognitive task. *Memory & Cognition*, *27*(1), 94–105.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232.

Milton, F., & Wills, A. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 407–415.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, *12*(4), 529–562.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*(4), 1035–1042.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., & McKinley, S. C. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology Section A*, *54*(1), 197–235.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*(1), 83–100.

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

Rafferty, A. N., & Griffiths, T. L. (2010). Optimal language learning: The importance of starting representative. In *Proceedings of the 32nd annual conference of the cognitive science society.*

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Reimers, S., Donkin, C., & Le Pelley, M. E. (2018). Perceptions of randomness in binary sequences: Normative, heuristic, or both? *Cognition*, *172*, 11–25.

Rogers, E. M. (2003). *Diffusion of innovations.* New York: NY: Free Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443-464.

Smith, D. J., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27.

Smith, S. M., Ward, T. B., & Finke, R. A. (1995). *The creative cognition approach.* Cambridge, MA: MIT Press.

Smith, S. M., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition*, *21*(6), 837–845.

Stewart, N., & Brown, G. D. A. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology*, *49*, 403–409.

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beston, G., ... McMullin, B. (2016). Open-ended evolution: Perspectives from the OEE workshop in York. *Artificial Life*, *22*(3), 408–423.

Tenenbaum, J. B., & Griffiths, T. (2001). The rational basis of representatives. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23).

Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(1), 119–143.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases.

Science, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*(1), 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* (pp. 157–178). Cambridge, MA: MIT Press.

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*(2), 199–216.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120.

A The Hierarchical Bayesian Model of Concept Generation

Jern and Kemp (2013) demonstrated how a hierarchical Bayesian model could explain the distributional correspondences between observed and generated categories. In their model, exemplars of generated categories were viewed as samples from a multivariate Normal distribution over the dimensions of stimulus space. The mean of the generated category was independent of the observed categories, but the covariance matrix (encoding feature variances and correlations) was based on a common prior distribution. Generating a new category was thus completed by sampling a new category mean (uniform over feature space) and covariance matrix from the common prior distribution. Because the shared prior distribution's parameters were unobserved, the hierarchical Bayesian approach was used to infer its parameters from the previous categories (their feature variances and correlations), and then to generate the covariance matrix of the new category.

In our implementation of their model[8], each category's exemplars are viewed as samples from a multivariate Normal distribution with parameters (μ, Σ). Category covariance matrices (specifying variance and covariance along k-dimensions), are assumed to be Normal-Inverse-Wishart distributed with parameters: ν $(> k-1)$, κ (> 0), and Σ_D. ν and κ are treated as free parameters in our simulations, and Σ_D is the domain-wide covariance matrix from which all categories are viewed as samples. Assuming a given Σ_D, a category covariance matrix Σ can be computed on the basis of its examples:

$$\Sigma = \left[\Sigma_D \nu + C + \frac{\kappa n}{\kappa + n}(\bar{x} - \mu)(\bar{x} - \mu)^T\right](\nu + n)^{-1} \tag{A.1}$$

where $\bar{x}$ and C are the empirical mean and covariance of the category's known members, and n is the number of observed members of the category. When there are fewer than two

[8] Note that Jern and Kemp (2013)'s model is slightly different, as they used a semi-conjugate model. Their model acts very similarly to our version.

known members of the category (and thus no covariance to speak of), $\Sigma = \Sigma_D \nu$.

The category mean, μ, can be computed as:

$$\mu = \frac{\kappa \mu_0 + n\bar{x}}{\kappa + n} \tag{A.2}$$

where μ_0 is the prior mean. In our simulations, μ_0 is set to the center of the domain. However, when no examples of the target category have been observed, generation is assumed to be random. In practice, the model's best fits are achieved when the κ parameter, which controls the influence of μ_0 on μ, is set very close to zero (hence, the influence of μ_0 is minimal).

Importantly, the domain-wide covariance matrix Σ_D is unobserved and needs to be inferred from the observed categories. For conjugacy, if Σ_D is viewed as a sample from an Inverse-Wishart distribution with scale Σ_0, Σ_D can be computed as:

$$\Sigma_D = \Sigma_0 + \sum_y C_y \tag{A.3}$$

where Σ_0 is the prior covariance in the domain. In our simulations, $\Sigma_0 = \lambda \mathbf{I}$, where λ is a free parameter controlling the expected variance of dimensions (dimensions of the domain covariance matrix are expected to be uncorrelated) and $\mathbf{I}$ is a k-by-k identity matrix.

Generated exemplars are drawn from a multivariate Normal distribution specified by (μ, Σ). Thus, $p(y)$ is

$$p(y \mid x) = \frac{\exp\{\theta \cdot \text{Normal}(y; \mu, \Sigma)\}}{\sum_i \exp\{\theta \cdot \text{Normal}(y_i; \mu, \Sigma)\}} \tag{A.4}$$

where θ is a response determinism parameter and $\text{Normal}(y; \mu, \Sigma)$ denotes a multivariate Normal density evaluated at y.

B Representativeness Model in Toroidal Space

Assuming a single bounded feature, we can easily represent the probability density of a function at any point x on a line as $f(x)$. With a single unbounded feature, the points x are mapped onto a unit circle with the corresponding $f(x)$ "wrapped" around the circle. More formally, we can represent the probability density along a wrapped axis $f'(x)$ as:

$$\begin{aligned} f'(x) &= ... + f(x-4\pi) + f(x-2\pi) + f(x) + f(x+2\pi) + f(x+4\pi) + ... &\text{(B.1)}\\ &= \sum_{i=-\infty}^{\infty} f(x+2\pi i) &\text{(B.2)}\end{aligned}$$

We implement the wrapped axis in the representativeness model when calculating the multivariate Gaussian components of the representativeness equation $R(x,h)$ (Equation 4 of the main text). Specifically, the probability density of the Gaussian component along a single wrapped axis can be computed as:

$$p'(x|h) = \sum_{i=-\infty}^{\infty} p(x+2\pi i|h) \qquad \text{(B.3)}$$

where $p(.|h)$ represents the probability density of the Gaussian component h on a line (i.e., not wrapped around a unit circle).

In practice, we can approximate the infinite summation by constraining i within a limited range (e.g., $[-50, 50]$). For increased efficiency, we can also approximate $p'(x|h)$ by stopping the summation once the ratio of the maximum and minimum values of $p(x+2\pi i|h)$ computed thus far is large enough (e.g., 10,000). For our modeling exercises here we impose both constraints, although the first constraint shows some redundancy since we find that i rarely, if ever, exceeds a value of 4. Given the unimodal nature of the Gaussian distribution, this procedure allows for comparatively efficient approximation of the wrapped distribution without needing to compute the infinite sum.