Creating Something Different: Similarity and Contrast in Concept Generation

Nolan Conaway[1], Kenneth J. Kurtz[2], and Joseph L. Austerweil[1]

[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: Joseph L. Austerweil, 1202 West Johnson Street, Madison, WI 53706. E-mail: austerweil@wisc.edu

Abstract

The ability to creatively 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 the degree of contrast between the known and generated categories as a factor. We report two experiments demonstrating that contrast between what is known and what is created is of fundamental importance for categorization and the creative process. We propose a novel, exemplar-based approach explaining our results, and compare the model's performance to two key alternatives. Our experiments and simulations demonstrate specifically how category contrast influences creative generation, and how it can interact with other constraints to produce different types of creations. Our work also serves as a concrete example of how well-established and highly controlled experimental frameworks (such as those in the field of category learning) can be used to make progress in an intriguing yet understudied domain.

Keywords: categorization; concepts; creativity; generation; computational modeling; exemplar models

1 Introduction

The creation of new ideas is one of the most fascinating and important human capabilities. Yet, it is also one of the most difficult human capabilities to study. We tend to focus on the most salient products of conceptual generation (e.g., scientific breakthroughs), but every person is likely to have generated many novel concepts in ordinary life. Indeed, the generation of novel sentences, works of art, and so on may also be considered products of our capacity to create new concepts. However, while this capacity is of great interest, it is also highly complex, making it difficult to study in a typical laboratory setting. By consequence, we do not have a strong understanding of the representations and processes involved in the generation of new concepts.

Questions concerning the generation of novel concepts are typically addressed within research on creativity, a tremendously diverse field employing a multitude of experimental and theoretical approaches (see **?**). The processes underlying conceptual generation are commonly investigated using the creative cognition approach (**??**), in which the products of creative acts (such as drawings) are analyzed to obtain insights into how they might have been created. These approaches have made great advances in our understanding of creativity, but the focus on particular types of creative products such as drawings makes it difficult to formalize the proposed processes and representations within a computational modeling framework. As a consequence, this approach does not often employ computational models to provide formal tests of competing theories.

The act of generating a new concept is, 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 (**??**). Research in categorization typically focuses on classification (predicting an object's category given its features; **??**) and inference (predicting an object's unobserved features given observed features and a category label, see e.g., **?**). 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 creative generation by extending theories of categorization to the special case.

Previous work in categorization 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 (**???**). 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 (**??**). Although this 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).

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. 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 ubiquitous throughout the categorization literature, and extends to a variety of other fields (e.g., color; **?**), 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. Building on the largely successful exemplar modeling framework (**???**), 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. This model makes novel predictions about how contrast affects category generation, which we test using behavioral experiments.

The outline of the article is as follows. First we describe previous empirical work on

the topic of category generation, as well as the computational formalizations of the theories in those reports. Then we describe our novel computational model, which is designed to generate categories that systematically differ from existing categories in the domain. We present two experiments demonstrating strong and systematic effects of category contrast on creative generation, and we qualitatively and quantitatively analyze the performance of each model in capturing human category generation. We conclude with a discussion of the implications of our results for categorization and creative cognition, and directions for future work.

2 Prior work

Much of what we know about concept generation comes from the foundational literature on creative cognition. In a series of reports, Ward and colleagues (**?????**) 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 as found in known categories. In a seminal study, **?** 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 nature of creative generation. They indicate that people rely strongly on prior knowledge in the creative process, and people generate concepts in accord with what they already know. Much of the work from this area (e.g., **??**) focuses on how information provided to participants (such as an example of a species generated by other participants) can drastically diminish creativity. Theoretical accounts of these effects have primarily

been grounded within the categorization literature. For example, the predominant ``Path of Least Resistance'' account (see **???**) 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 (**??**), as well as the exemplar view (**??**). While this work is been incredibly useful in providing a conceptual sketch of generation theories, the hand-drawn responses used in the experiments paradigms precludes the development of formal approaches that can be used to test the theories in a fine-grained manner.

? recently showed that creative generation could be studied in a more controlled manner through the well-developed methods of an artificial categorization paradigm (see **?**, 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 that of the **?** studies, **?** found that participants generated categories with the same distributional properties as the experimenter-defined categories. For example, after training on 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. This finding is notable, as it demonstrates 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.

? 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 in which features are changed (both of which are factors in the Path of Least Resistance account; **?**). Instead, their copy-and-tweak model corresponds to a direct exemplar-similarity approach (e.g., **??**), 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 **?** were specifically designed to challenge it. However, its application is notable as a first step toward understanding theories developed in the creative cognition literature using well-known formal approaches from the categorization literature.

3 And Now for Something Different: A Role For Contrast

Prior work on category generation has explored only one of the possible constraints that guide category generation. The main concern of the published experiments has been on the distributional correspondences between learned and generated categories, and as a result most of the computational, theoretical, and empirical efforts have been focused on explaining those effects. In this paper, we investigate another important constraint: category contrast. To generate a novel concept, individuals must produce something that is

in some capacity *different* from what they already know. By consequence, we propose that contrast should be a primary constraint on creative generation: New concepts should be different from existing ones.

Although it is evident that people are *capable* of creating new concepts and categories, it is not entirely clear how new concepts are systematically made different from what is already known. The hierarchical Bayesian model developed by **?** 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 with a broader class of standard hierarchical Bayesian models (e.g., **??**). 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*. The best a standard hierarchical Bayesian model can do at capturing contrast is to assume that the new category is placed uniformly at random over stimulus space. But, this defeats the purpose of a hierarchy as it is ignored when determining a new category location![1]

The copy-and-tweak model tested by **?** 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

[1]It is plausible that some hierarchical Bayesian model could be created that generates 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 promotes contrast.

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.

3.1 Novel Analyses Demonstrating Contrast Effects in Prior Work

Although existing accounts of creative generation broadly overlook the role of category contrast in determining what is novel versus familiar, it was implicitly assumed that learners in previous experiments were *successful* in creating new categories. Thus, effects of category contrast should be observable within the experimental results of these studies. To provide a test of the influence of category contrast within existing data, we conducted a novel analysis of Experiment 3 from **?**[2].

Participants in their experiment were exposed to members of two experimenter-defined categories of `crystals' varying in hue, saturation, and size. Each category possessed a unique hue, but varied in saturation and size: In the `Positive' condition, there was a positive correlation between these features (i.e., larger sized crystals were more saturated), and in the `Negative' condition this relation was reversed. In the `Neutral' condition, there was no correlation between saturation and size. After learning about the categories from each condition, participants were asked to generate six exemplars belonging to a novel class. As noted above, **?** found that the generated categories tended to follow the distributional properties of the experimenter-defined categories: Generated categories were tightly distributed along the hue feature, and possessed the same saturation-size correlations as in the learned categories. **?**, however, did not analyze or discuss how the generated categories *differed* from the experimenter-defined categories.

Because each experimenter-defined category in the **?** experiment possessed a

[2]Although **?** reported four experiments, we focus on Experiment 3 because their first two experiments tested generation of items into known categories, and their fourth experiment was identical to Experiment 3 but with a restricted generation space.

distinct hue shared by all members of the category, it is sensible that participants might generate a category with a hue distinct from the experimenter-provided categories. If creative generation were influenced by category contrast in this way, the hues of generated categories should be systematically different from those of the experimenter-defined categories. Unfortunately, stimulus hue was encoded and presented in the Hue-Saturation-Value (HSV) color space, which is device-dependent and not perceptually normed such that perceived color similarity corresponds to proximity in the color space (as opposed to a color space such as CIELAB that is device-independent and equidistant sets of points correspond to pairs of colors that have the same perceptual similarity; **?**). Further, they did not calibrate their monitor, and so we cannot know the precise colors presented to participants. As **?** were interested in relations between the saturation and length of examples in generated into novel categories, these issues do not undercut their analyses and results. However, these issues pose a significant challenge to evaluating contrast between the experimenter-defined and participant-generated categories along the hue dimension. It is plausible that two uncalibrated monitors could display the same HSV color and the colors be perceived in different color categories (especially for color boundaries that vary over lightness, such as the yellow-brown boundary).

Although we cannot know the precise colors that were displayed or perceived, we can still analyze their results from a coarse perspective to see whether there is preliminary support for contrast. To do so, we binned all possible hues into one of eight uniformly-spaced color groups: $\{$*Red*: $0 - 0.063$, $0.938 - 1$, *Yellow*: $0.063 - 0.188$, *Yellow-Green*: $0.188 - 0.313$, *Green-Teal*: $0.313 - 0.438$, *Teal*: $0.438 - 0.563$, *Teal-Blue*: $0.563 - 0.688$, *Purple*: $0.688 - 0.813$, *Pink*: $0.813 - 0.938\}$. In the **?** experiment, the hue of each experimenter-defined category was selected from one of six possible values, each of which falls into one of the color groups above (two color groups were not used as a possible hue for the experimenter-defined categories). By categorizing the participant-generated crystals likewise, we can obtain a broad measure of category contrast by determining the

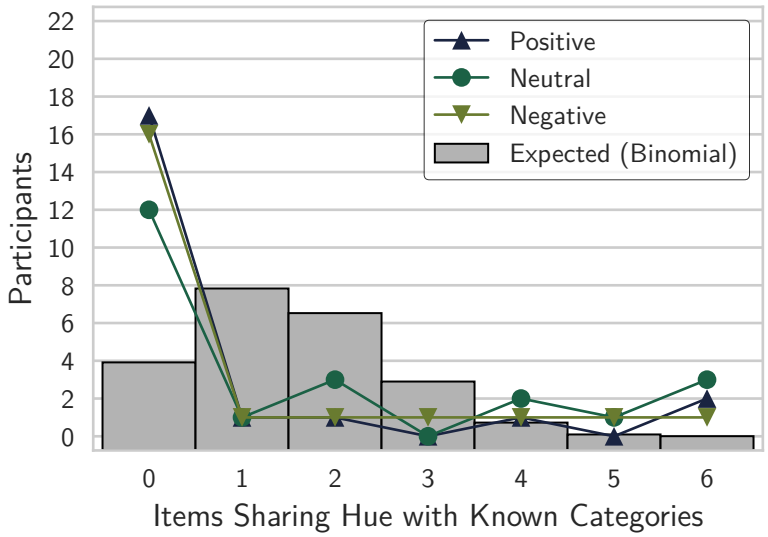

Figure 1: Analysis of data from **?**, Experiment 3. Plotted is the number of generated items that share a color group with one of the experimenter-defined classes. The ``Expected'' data follows a Binomial distribution with $p = 2/8 = 1/4$, given there were two experimenter-defined classes, and eight color groups.

proportion of participant-generated crystals that fall into the same groups as the experimenter-defined categories: If contrast influences the hues of the generated categories, we should observe minimal overlap between in the color groupings.

These data, shown in Figure 1, reveal a clear pattern: The majority of participants in each condition ($n = 22$) generated categories possessing entirely distinct hues; with 0/6 exemplars sharing a hue with the experimenter-defined categories. These results can be compared to the predictions of a Binomial model, which proposes that participants generate hues at random. That is, if hue selection is not systematic, the probability that any given example will lie in the same color group as an experimenter-defined category is given by a Binomial distribution with $p = 2/8 = 1/4$, as there were two experimenter-defined categories and eight possible color groups. Chi-square goodness-of-fit tests reveal that the observed distribution in each condition is highly inconsistent with the hues being chosen at random, all $\chi^2(6, N = 22) > 200$, $p < 0.001$. Participants tended to generate items that were perceptually distinct from the categories they had learned, and were less likely to

generate hues possessed by members of the experimenter-defined categories.

Re-analyzing the results from **?** provides preliminary support that contrast plays a role in category generation. Taken alongside the analyses reported by **?**, our analysis suggests that generated categories tend to be distinct from *and* distributionally similar to what is already known. However, it is worth noting that our analysis is still limited: The color groups defined above are imprecise, and it is not clear that our color grouping is consistent with the psychological color boundaries perceived by participants. While we did obtain similar results using a variety of alternative groupings, the hue dimension used in the **?** study does not lend itself straightforwardly to the the computation of similarities, and thus we cannot be certain of whether our coding accurately approximates the psychological space of the stimuli. This precludes traditional applications of categorization models to their data as it is usually necessary to encode objects in psychological space in order to accurately determine the similarity between objects. By consequence, although these results likely indicate that contrast exerts *some* influence, they do not precisely describe the nature of that influence. In the sections below, we propose a quantitative framework specifying the role of category contrast in creative generation.

3.2 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 creative 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 fundamental constraints: 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; **??**).

Although, as an exemplar model, one of PACKER's proposals is people represent categories in terms of a collection of stored exemplars, it is worth emphasizing that the underlying nature of human category representations is not the primary concern of our work. The choice to develop PACKER within an exemplar framework reflects the facts that exemplar models have been throughly evaluated, are strongly theoretically motivated, and dominate much of the theoretical and empirical work in categorization. The focus of our work with PACKER concerns 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., **???**; for a review of categorization models, see **?**).

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 **???**):

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

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, attention will be distributed uniformly in our simulations (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., **??**). 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. Specifically, **?** 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: $f(x_j) = \gamma$ if x_j is a member of the target category, and $f(x_j) = \gamma - 1$ if x_j is a member of a contrast category. γ is thus a free parameter ($0 \leq \gamma \leq 1$) controlling the trade-off between within- and between-category similarity. For example, when $\gamma = 0.5$, $f(x_j) = 0.5$ 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, $\gamma = 1$ produces exclusive consideration of target-category members, and $\gamma = 0$ produces exclusive consideration of opposite-category members. The γ parameter thus specifies a wide breadth of possible

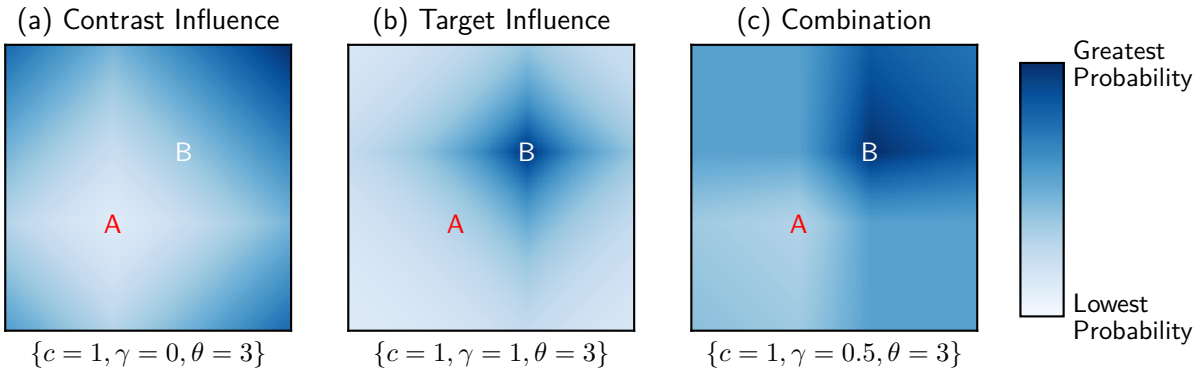

Figure 2: PACKER generation of a category `B' example, following exposure to one member of category `A' and one member of category `B'. Predictions are shown for three different parameterizations (differing only in γ): *(a)* Predictions based on contrast similarity only. *(b)* Predictions based on target similarity only. *(c)* Predictions with both constraints considered.

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 2 for an illustration of how γ controls 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 **?** choice rule. Candidates with greater values of $a(y, x)$ are more likely to be generated than candidates with smaller values:

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

where θ ($\theta \geq 0$) is a free parameter controlling response determinism.

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, as will be discussed in more detail below, PACKER is formally equivalent to the copy-and-tweak model proposed by **?** when $\gamma = 1$. Likewise,

when $\gamma = 0$, PACKER can represent a contrast-only generation mode, relying exclusively on contrast when generating new categories. Finally, when $f(x_j) = -1$ 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.

It is also worth noting that PACKER represents just one of many possible models that incorporate contrast in category generation. For example, it may be possible to extend **?**'s hierarchical Bayesian model to include contrast, but this mechanism would be at odds with the standard hierarchical Bayesian framework. In contrast, these ideas emerge naturally from the exemplar view: We have not modified any of the core elements of the GCM in defining the PACKER model, we simply weight exemplar similarity during aggregation. Thus, beyond formally specifying the role of contrast in creative generation, the PACKER account allows us to evaluate the well-understood principles of the exemplar view within the comparatively understudied field of conceptual generation. Nonetheless, we will discuss integrating our approaches in the General Discussion.

3.2.1 Relation Between PACKER and Copy-And-Tweak

The PACKER model is similar to the copy-and-tweak model reported by **?**: Both models are exemplar-based, and both models generate new items according to their similarity to known members of the target class. However, PACKER diverges from the copy-and-tweak model by including a contrast mechanism, enabling generation according to dissimilarity to members of opposing categories. By consequence, copy-and-tweak can be realized as a parameterization of the PACKER model that is insensitive to category contrast. Specifically, when $\gamma = 1$ (see Figure 2, panel B), $f(x_j) = 0$ for x_j belonging to contrast categories; thus, PACKER is not influenced by these items, and is mathematically equivalent to a copy-and-tweak approach.

In this paper, we report simulations using this copy-and-tweak model. This model

fits within the exemplar-based category generation framework defined above, under the constraint that $\gamma = 1$, and is a continuous-dimension adaptation of the model tested by **?**. 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: 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 $\gamma < 1$ significantly improves model fits). Note that the purpose of the article is to explore and formally analyze the role of contrast in category generation and thus, we leave extending PACKER to incorporate distributional factors (as explored by **?**) for future work.

3.3 Synopsis and Prognosis

Research on the creative generation of novel concepts has focused on the finding that generated categories tend to possess distributional commonalities with known categories. However, a fundamental goal of concept generation is to create something *new* (i.e., different from what is already known). The manner in which generated categories differ from known ones is, nonetheless, poorly understood: Existing theories do not make strong predictions about how creatively generated concepts should systematically differ from existing ones. Above, we found encouraging initial support that contrast influences category generation (**?**, Experiment 3), and we introduced a novel, exemplar-based model formalizing the roles of similarity and contrast in creative generation.

In the sections below, we present two experiments demonstrating systematic effects of category contrast on creative generation inspired by factors influencing how PACKER generates new categories. Our experiments are based on **?**'s paradigm, which is a straightforward translation of the traditional artificial classification paradigm to the task of category generation: Participants are first exposed to a single, experimenter-defined category, and are then asked to generate members of a new category. We then report

formal analyses comparing PACKER's account of our results to that of the hierarchical Bayesian and copy-and-tweak models developed by **?**.

4 Experiment 1

To begin our investigation, we sought to extend the early evidence obtained from our analysis of the **?** data, under a variety of learning conditions and using more standard stimulus materials. We used an artificial stimulus design: A two dimensional domain of squares, varying in color and size (see Figure 3, panel A). These dimensions have been used in numerous classification learning studies (e.g., **????**). Unlike those used in the **?** experiments, distance on these physical dimensions aligns more directly with perceptual similarity, allowing us to evaluate the role of category contrast in creative generation more precisely. To extend the evidence provided by the **?** data, 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 between the perceptual and conceptual dimensions (e.g., $X \rightarrow Size$, $Y \rightarrow Color$), as well as the direction of

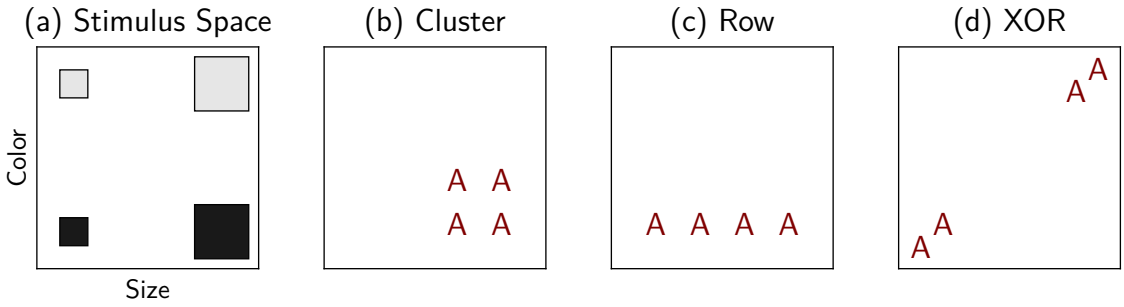

Figure 3: Stimulus domain and category types tested in Experiment 1. Stimuli are not drawn to scale.

variation along each dimension (e.g., *dark* → *light* or *light* → *dark*) was counterbalanced across participants. The category types in Figure 3 are plotted in a conceptual space, rather than a 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 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 of the classic finding that

generated categories tend to share distributional properties with known categories in the domain (**??**). 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 be negatively correlated instead. In this case, contrast and statistical regularities would interact, which would be inconsistent with the leading approach in conceptual generation (**?**)

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 domain space (x, y), as well as the direction of variation along each axis (e.g., $dark \rightarrow light$ or $light \rightarrow dark$) was counterbalanced across participants.

4.2 Procedure

As noted in the introduction of this paper, the task of generating members of a new category can be aligned with common tasks studied 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. In keeping with this insight, 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., **?**), 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 **?**, 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.
>
> 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 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

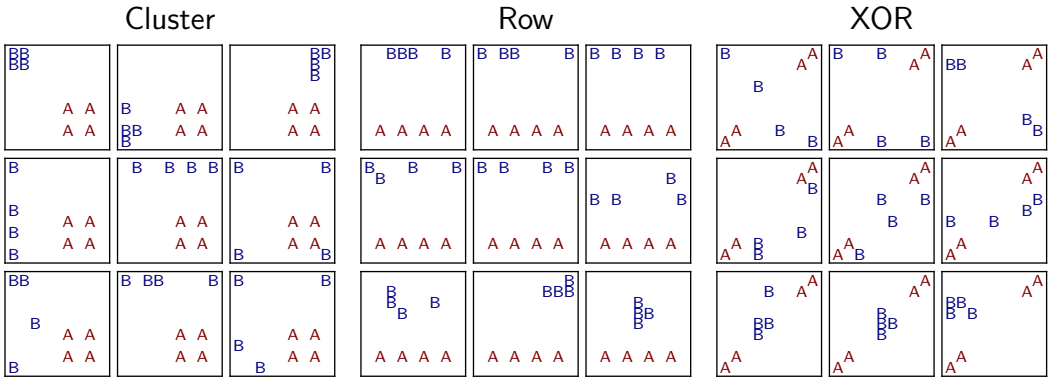

Figure 4: Sample categories generated by participants in Experiment 1. Representative samples from common generation profiles are shown.

that contrast is a fundamental constraint on creative cognition and that suggests that statistical regularity alone is insufficient.

Figure 5 also depicts, for each participant, the average distance of members within 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, $t(63) = 11.43$, $p < 0.001$; Row, $t(60) = 13.16$, $p < 0.001$; and XOR, $t(57) = 3.64$, $p < 0.001$. 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.

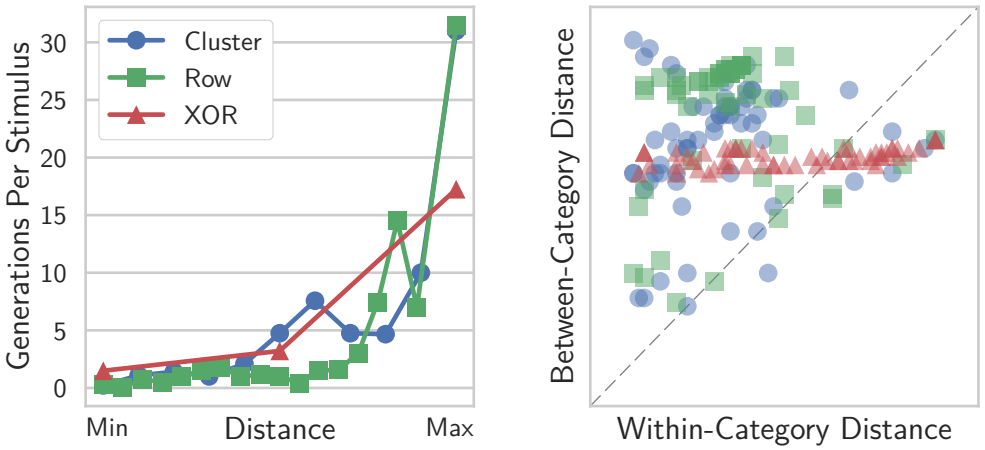

Figure 5: Experiment 1 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

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. 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, participants generated categories spanning the entire X- and Y- axis as well as 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, $t(123) = 5.61$, $p < 0.001$, and XOR, $t(120) = 2.68$, $p = 0.008$. Categories generated in the XOR condition were also less widely distributed along the X-axis compared to Row, $t(117) = 2.56$, $p = 0.012$. This latter effect was not expected because the

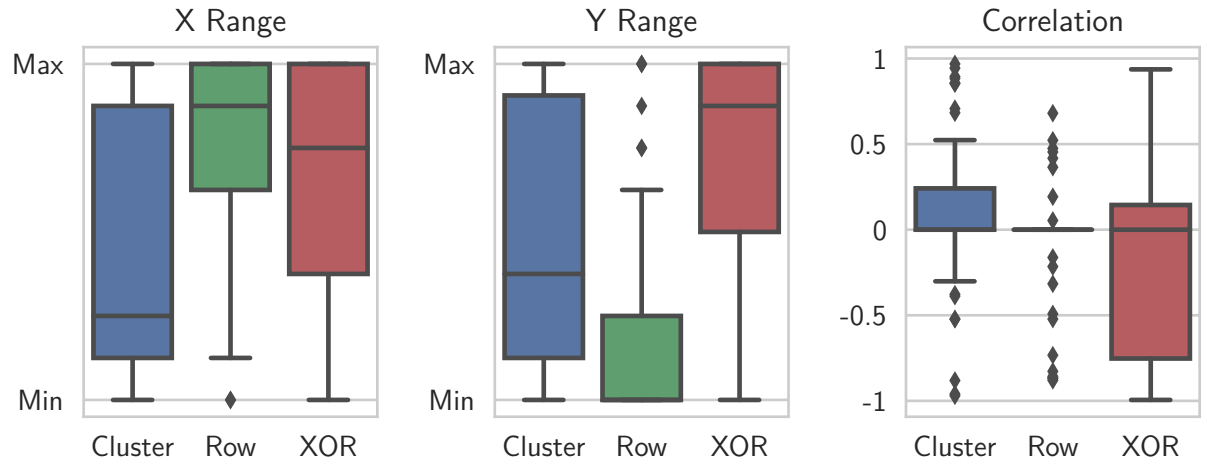

Figure 6: Box-plots of the distributional statistics from the categories generated in Experiment 1. Boxes depict the median and quartiles of each condition, with whiskers placed at 1.5 IQR. All points outside this region are marked individually.

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, $t(123) = 4.57$, $p < 0.001$ and XOR, $t(117) = 9.26$, $p < 0.001$, and categories from the Cluster condition had less Y-axis range compared to XOR, $t(120) = 3.95$, $p < 0.001$. As expected, the correlations in the Cluster and Row conditions were not systematically positive or negative ($ps > 0.1$). However, the generated categories in the XOR condition tended to possess *negatively* correlated dimensions, $t(57) = 2.04$, $p = 0.046$. This finding is notable, as it is the opposite of what would be expected, 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 this finding is because participants in **?** could differentiate the generated category on a third dimension (hue) to maintain the statistical regularities on the other two dimensions. 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 category generation.

4.4 Discussion

In Experiment 1 we sought to extend our analysis of the **?** data by evaluating the influence of category contrast on creative generation, given qualitatively different types of prior knowledge. 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 (**??**): 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

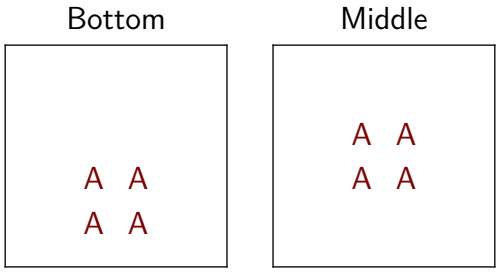

Figure 7: Category types tested in Experiment 2.

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 constraints on creative 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 in the bottom-center 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 stimulus 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 stimulus 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.

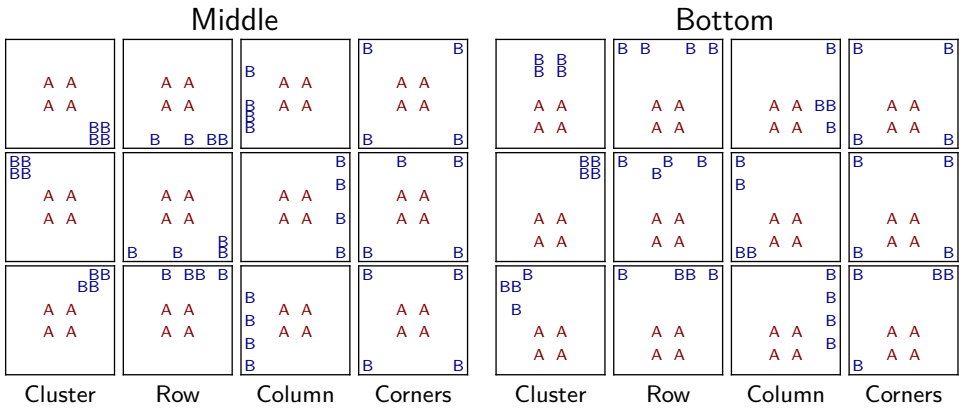

Figure 8: Sample categories generated in Experiment 2.

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 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, $t(60) = 5.5$, $p < 0.001$; Middle, $t(60) = 2.71$, $p < 0.01$. We did, however, observe a notable

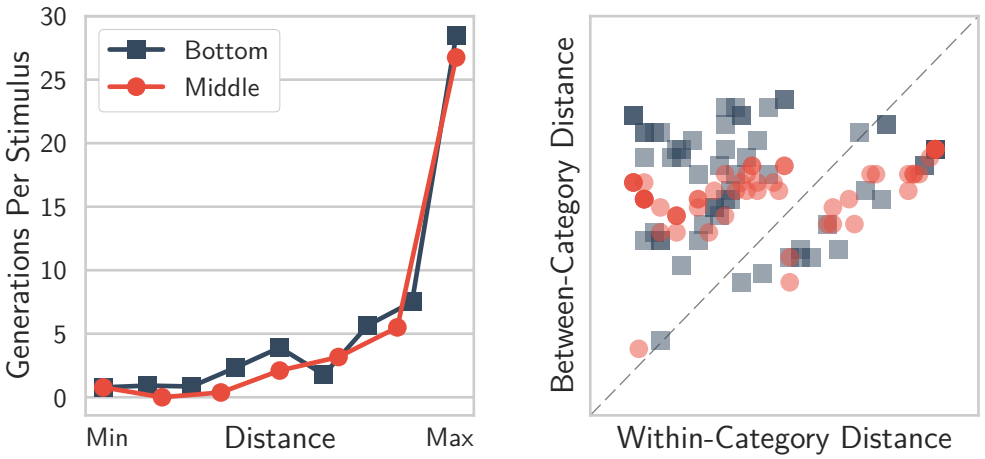

Figure 9: Experiment 2 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

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 observed in our study. In each condition, we observed tightly clustered and widely distributed categories along each dimension, as well as positively, negatively, and uncorrelated 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

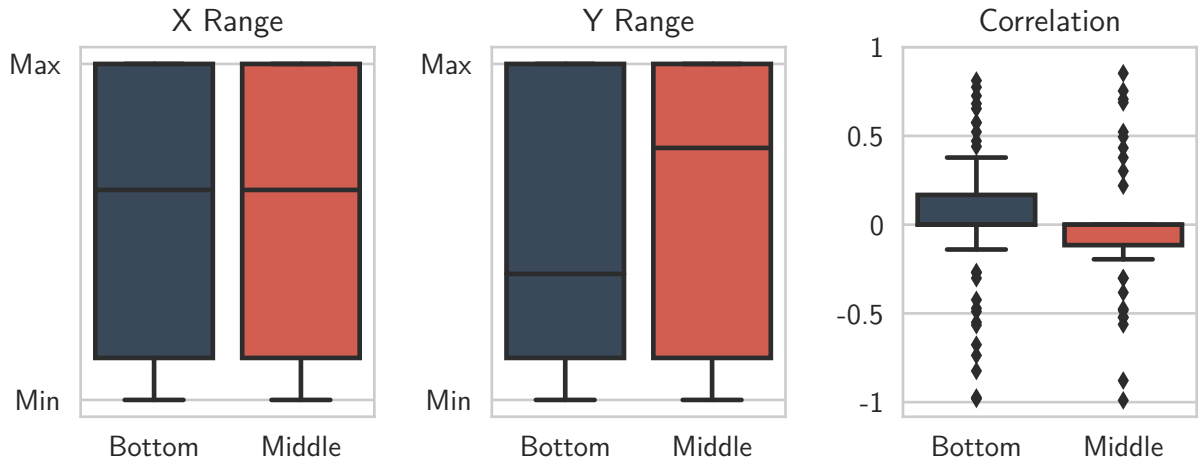

Figure 10: Box-plots of the distributional statistics from the categories generated in Experiment 2.

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 ($ps > 0.17$).

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 $W = 0.77, p < 0.001$ for the Middle condition and $W = 0.85, p < 0.001$ for the Bottom condition). Figure 11 depicts the Y-axis position of the exemplars generated within each participant's category. The categories are sorted by overall range, then by condition assignment. 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. The non-normality of the Y-axis range distributions thus requires that we use a different approach to addressing the experiment's main question.

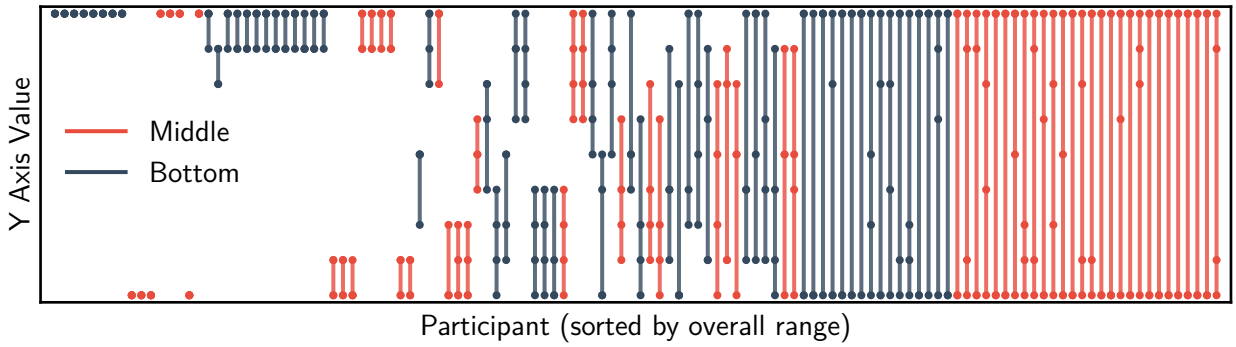

Figure 11: Y-Axis range and position of the participant-generated categories from Experiment 2. Each line corresponds to a participant's category, with notches corresponding to the Y-axis position of exemplars within the category (notches may overlap). Participants are sorted by overall range, and then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

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' 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, $p < 0.001$, 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, $p = 0.16$, however, more Middle participants placed exemplars in the top *and* bottom rows, $p = 0.038$. The latter effect is of 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, distributional structure of the generated categories can be influenced 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

Experiments 1 and 2 revealed systematic and strong effects of category contrast on creative generation. In this section, we report the results of simulations with formal models aimed at explaining our observations. Specifically, we present simulations from the PACKER model, as well as a `copy-and-tweak' model (discussed in Section 3.2.1), defined as a variant of PACKER with the γ parameter constrained to be one. The comparison of these two models serves to highlight the explanatory role of contrast within PACKER's framework: If contrast affords little explanatory advantage, then the two accounts should produce an equally strong account. We also present simulations from an implementation of the hierarchical Bayesian model proposed by **?**, described in-depth in Appendix A. The comparison between the hierarchical Bayesian model and PACKER is meant to 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 is insensitive to the distributional structure of learned categories (relying only on within- and between-category similarity), the 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**	**Hierarchical Bayesian**
$AIC = 9095$	$AIC = 9842$	$AIC = 9912$
$L = -4545$	$L = -4919$	$L = -4952$
$c = 0.482$	$c = 3.187$	$\kappa < 0.001$
$\gamma = 0.525$	$\gamma = 1$ (fixed)	$\nu = 5.596$
$\theta = 6.664$	$\theta = 2.969$	$\lambda = 0.055$
		$\theta = 3.174$

6.1 Parameter-Fitting

To obtain a global measure of the quality of each model's account, we fitted 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 fitted three parameters in the PACKER model (c, γ, and θ; see Section 3.2), as well as four in the hierarchical Bayesian model (κ, λ, ν, and θ; see Appendix A). We fitted only two parameters for the copy-and-tweak model (c, and θ), as γ is held constant ($\gamma = 1$). Note that each model possesses a θ parameter fulfilling the same role (response determinism). 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 results of this fitting procedure. Due to the uneven number of fitted parameters among the models, we compare the model fits using the Akaike Information Criterion (AIC; **?**), 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, **?**; AIC_C, **?**).

Table 2 contains the AIC values of each model's best fit, as well as the corresponding log-likelihood (L) and the best-fitting parameter values. These results reveal strong model differentiation: The PACKER model achieved far better fits compared to the

copy-and-tweak and hierarchical Bayesian models, and copy-and-tweak performed somewhat better than the hierarchical Bayesian model. While PACKER's advantage 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, as opposed to the prototype-based representation 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 **?** model, but can be easily accounted for using an exemplar-based approach.

A key distinction between PACKER and copy-and-tweak, as well as the hierarchical Bayesian model, is that, of the three models, only PACKER is 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 these 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 this trial (leaving 915 responses in the dataset): Again, PACKER ($L = -3390$, $AIC = 6786$) achieved better fits than the copy-and-tweak ($L = -3579$, $AIC = 7162$) and hierarchical Bayesian ($L = -3612$, $AIC = 7232$) 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 ($\chi^2(1) = 749, p < 0.001$ for all data and $\chi^2(1) = 377, p < 0.001$ excluding the first example), providing further evidence that category generation is better

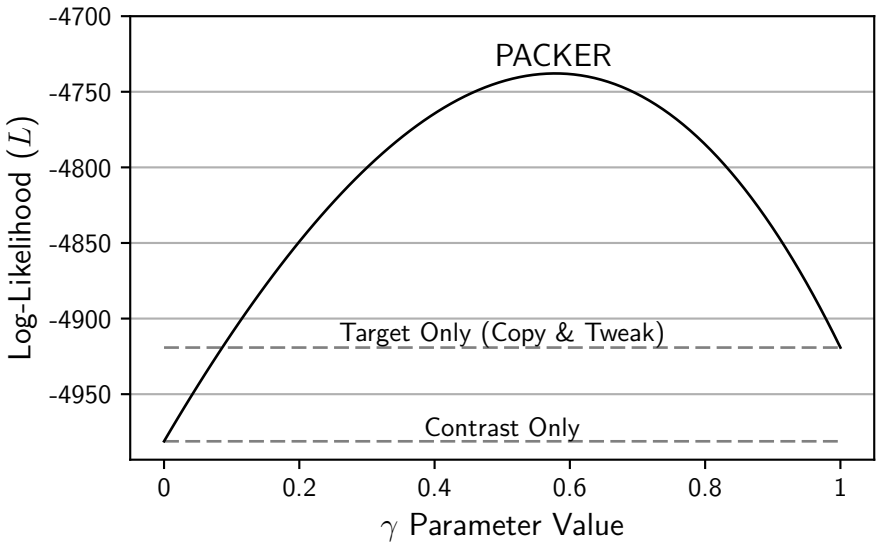

Figure 12: PACKER's fit as a function of its prioritization of within-category and between-category similarity (using the γ parameter). To facilitate comparison, PACKER's other parameters (c, θ) were set to the best fitting values obtained for copy-and-tweak in Table 2.

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 γ parameter. The model's other parameters (c, θ) were set according to copy-and-tweak's best fits from Table 2, and thus when $\gamma = 1$, 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 γ. The best fitting values lie well below the value of 1 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). Notably, however, the copy-and-tweak parameterization ($\gamma = 1$) performs better than the `contrast-only' parameterization of the model ($\gamma = 0$). 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 we observed can be described within the PACKER framework, 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

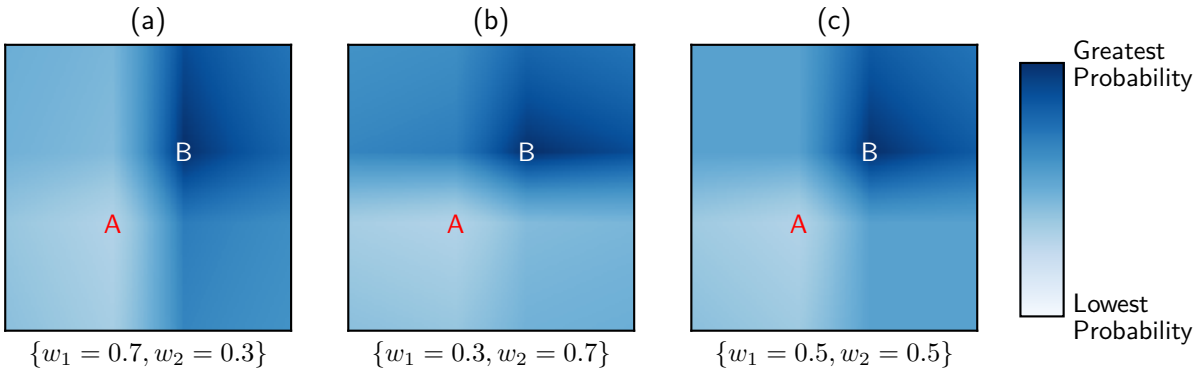

Figure 13: PACKER generation of a category `B' example, following exposure to one member of category `A' and one member of category `B'. Predictions are shown for different attention settings: *(a)* Increased weighting of the X-axis. *(b)* Increased weighting of the Y-axis. *(c)* Uniform weighting (identical to Figure 2).

(i.e., γ near 0).

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 (**???**), 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.

Similar to the copy-and-tweak model, the hierarchical Bayesian model possesses no mechanism to account for category contrast, and the model is most likely to generate new items that are similar to known examples of the target category. Although this precludes an account of why individuals might assume a cluster versus corners profile, the covariance matrix specifying the model's prior domain distribution, 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 dimensional correlations) 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.

As noted above, 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 creative 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:

$$\mu_x = \frac{\nu_0 \mu_0 + \sum x}{\nu_0 + n} \tag{4}$$

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[3]. 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

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

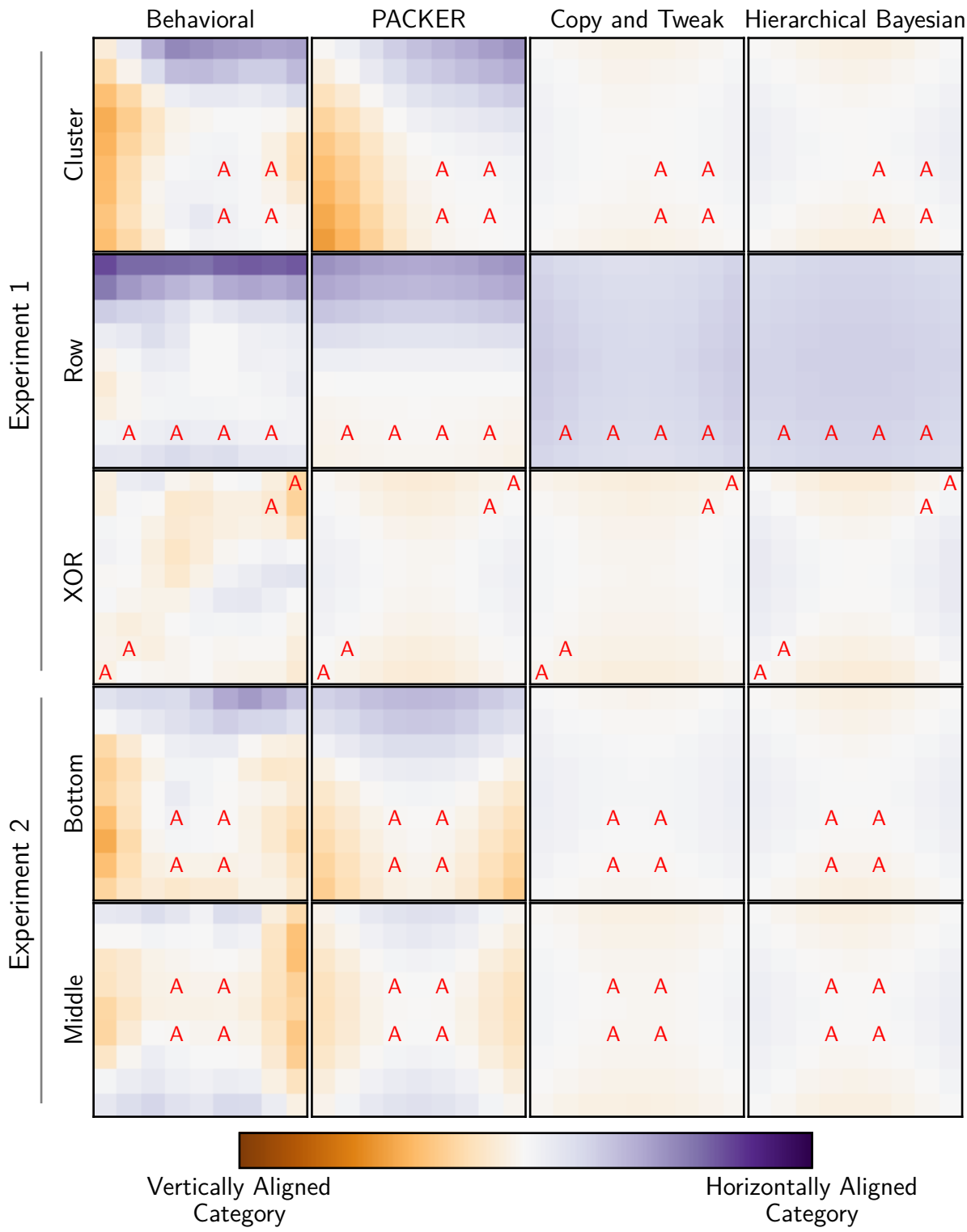

Figure 14: Behavioral and simulated range difference gradients. Each panel shows, for each stimulus, the dimensional orientation of the categories it was generated into: vertically aligned `columns' (orange) versus horizontally aligned `rows' (purple).

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 degree of within-category similarity.

To explore how well the PACKER, copy-and-tweak, 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 **?**), for simplicity we may assume that the Alpha and Beta categories are distinct along dimensions that the Betas do not vary on. Thus, the weighting for a given participant can be computed as:

$$w_k = \frac{\exp\{-\theta_w \cdot \text{range}(k)\}}{\sum_k \exp\{-\theta_w \cdot \text{range}(k)\}} \tag{5}$$

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 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 5. 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[4]. 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, PACKER provided a much closer match to our behavioral results than the copy-and-tweak and hierarchical Bayesian models. In all conditions, PACKER distributes 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

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

generated across the entire domain, rather than concentrated within the upper-regions as observed behaviorally.

7 Experiment 3.

[I'm assuming Joe will include some preamble for the experiment, but otherwise it should be fairly straightforward for me to churn out one anyway.]

7.1 Participants, Materials, and Procedure

Experiment 2 recruited 122 participants who each generated one set of four Alpha and four Beta category exemplars. Among these 122 generated category sets were 102 unique sets (i.e., they contained a unique collection of Alpha and Beta exemplars). Consequently, for Experiment 3, we recruited 102 participants with one participant presented with a different unique category set.

Participants observed four blocks of eight trials. Each trial began with the presentation of a fixation cross for 500 ms. This was followed by the presentation of one exemplar randomly sampled without replacement from the unique category set. Participants were tasked with assigning the presented exemplar to either the Alpha or Beta category with no time limit imposed. Feedback was automatically displayed for 2500 ms after each response.

7.2 Results

Overall accuracy of the participants was high, with a mean error rate of .19 ($SD = .19$). Error rates for each block are presented in Figure 1.

In the previous section (or Section X.X), we demonstrated that the best performing models were the two contrast models. In this section, our goal is to investigate if contrast is also important in category learning tasks.

It is also the goal of this paper to demonstrate that there is an association between the model's [justifying the errorvscat plot is harder than I thought] To compare the performances of each model, we analysed the correlation between each model's fit to a

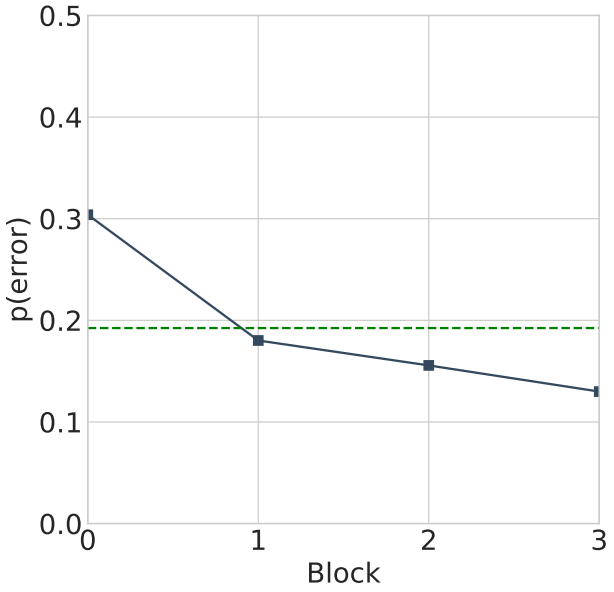

Figure 15: Average error rate for each successive block. Green discontinuous line represents the overall mean error rate.

participant's unique category set and the participant's error rate. Intuitively, a well-performing model should be able to closely fit (i.e., easily generate) A positive correlation would indicate that the model is We used the same set of optimised parameters found from Section X.X, with individual values for selective attention.

To emphasize the strong influence of contrast in maximizing the association between category generation and category learning, we computed the correlations over wide range of $\theta_contrast$ values, with the other parameter values of PACKER held constant at their optimized levels. As presented in Figure **??**, the correlation quickly increases with increasing weight on contrast. [more explanation pls]

The specific reason for why contrast as implemented in PACKER is superior to contrast as implemented in the representativeness model is still not clear.

[Note Joe's comments on figure formatting (font size, remove redundant y-axes, etc)]

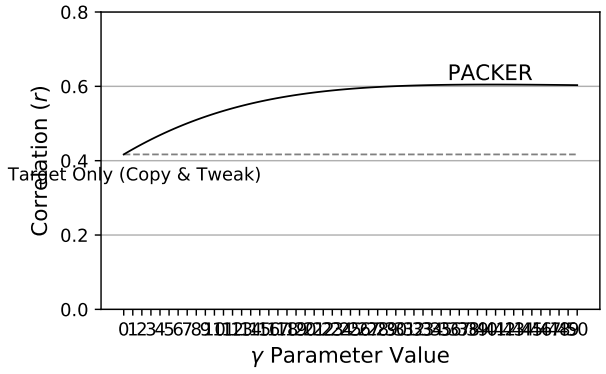

Figure 16: Correlation between PACKER's fit and participant error as a function of the $\theta_contrast$ parameter. To facilitate comparison, PACKER's other parameters (c, θ_target) were set to the best fitting values obtained for copy-and-tweak in Table 2.

8 General Discussion

The creative generation of concepts is an immensely intriguing topic, but it is also complex and difficult to study in a traditional laboratory setting. Creative generation is most typically studied using a creative cognition approach (**??**), where creative products (such as drawings of alien plants and animals; **?**) are analyzed to obtain insight into the processes and representations involved. While this approach has yielded a great deal of information about the nature of creative generation, the commonly employed response modalities (e.g., hand-drawn pictures) make it difficult to develop computational models capable of formally evaluating the proposed representations and processes. However, the generation of novel concepts is not altogether different from other behaviors studied in cognitive psychology: In particular, it can be viewed as an form of inference using one's knowledge about existing categories (**??**). Taking advantage of this insight, in this paper we investigated the generation of new categories using the empirical and computational toolkit from the field of categorization.

While the bulk of prior research on the topic 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 creative generation: How is it that people are able to create something *different* from what is already known? We developed a novel, exemplar-based model, PACKER, which formally specifies the role of contrast in generation. 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.

In addition to an analysis of published data (**?**, Experiment 3), we reported two experiments demonstrating systematic effects of category contrast in creative 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 stimulus 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 PACKER's account of our results to the other major 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 PACKER's sensitivity to contrast captured a previously unexplained and unexplored aspect of human category generation. Further, 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 creative generation.

8.1 Similarity and Contrast in Cognition

The proposal that category contrast is a primary constraint in creative generation is, in some ways, entirely commonsense. To successful create something *new*, it must be different from what is known. Beyond its role in creative generation, 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 (**??**). Likewise, basic-level categories (**?**) 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 (**??**).

Beyond categorization, one can find instances of the trade-off between within and between-class similarity in linguistic categories over perceptual dimensions. For example, **?** 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 (**??**). 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 successfully interprets this trade-off within the domain of creative generation, and allows us to begin to understand the relatively understudied processes involved in creative generation through our more well-developed knowledge of human categorization.

8.2 Implications for Creative Cognition

Although the focus of this article has been to address the processes involved in creative generation using the empirical and quantitative toolkit of traditional categorization research, 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 (**??**). 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 in the study creative cognition, the PACKER model more specifically offers 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 (**??**). 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 with PACKER demonstrate that this influence is concisely explained in terms of a trade-off between within-category similarity and between-category dissimilarity.

Similarly, 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 (**??**), 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 **?**), 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.

8.3 Implications For Categorization

Although our work has mainly addressed the processes involved in the creative generation of concepts, in this paper we have studied these processes as they apply within an artificial categorization domain. By consequence, our findings also provide an advance in our understanding categorization more broadly.

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 single category through positive examples and then figure out another category that would make sense in the domain. In other words, only one category is provided to be learned and the use that is required of that category is to generate a contrast category. Such propagating of categories (producing one from another) may not seem like a fundamental or ubiquitous cognitive activity, but consider navigating a highly unfamiliar environment -- it would be practical, perhaps even critical, to generate expectations about what other kinds of things you are likely to encounter based on those that you already have. In the present studies, we have learned something about the form that such expectations 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). In this sense one can see the patterns of performance in the category generation task as recapitulating the order of semantic organization.

8.4 Limitations and Future Directions

Although successful in explaining our results, 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 **??**). 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's success over the hierarchical Bayesian model indicates that the emulation of distributional structure exerts only a limited influence in comparison to category contrast in some situations.

Nonetheless, these classic effects are a core element of the phenomenology of creative 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 **?** with PACKER's underlying claims to obtain a ``best of both worlds'' model, capable of explaining the role of contrast in creative 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 **??**).

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 (**?**), 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. Alternatively, it may be possible to integrate the core principles of PACKER into other categorization models (e.g., **???**).

Finally, although an overall goal of this work is to advance the study of creative thought by explaining generation through empirical and quantitative approaches from research in categorization, these approaches are not easily aligned with many of the alternative theoretical viewpoints in the broader study of creativity (for a review see **?**), such as those based on free association (**?**) and conceptual combination (**??**). Instead, we reduced our focus by studying a highly complex behavior (creative 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.

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 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. More generally, our results demonstrate that popular explanatory approaches from basic research in cognitive science can offer a precise, quantitative account of a behavior as complex as that of creative generation.

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. We thank Alan Jern and Charles Kemp for providing code and data.

A The Hierarchical Bayesian Model of Concept Generation

? 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 stimulus 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[5], 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

[5]Note that **?**'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.