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

Abstract

NBC: Make sure to write this!

Keywords: categorization, concepts, creativity, generation

1 Introduction

Creativity, innovation, imagination, and the creation of new ideas are among the most fascinating, important and difficult human capabilities to study. They are fascinating and important capabilities because they have produced some of the greatest scientific breakthroughs that revolutionized the world. Unfortunately, technological and scientific breakthroughs are rare and the result of a great deal of cognitive effort, which makes it difficult to investigate the underlying cognitive representations and processes responsible for them using standard experimental methodology or to formalize in computational models. Although we tend to focus on the most salient products of these processes (e.g., scientific breakthroughs), every person is likely to have generated many novel sentences, thoughts, and/or drawings given the combinatorial explosion of possible statements given a reasonably-sized set of primitive elements and combination rules in language, thought, and perception (Goldstone, 2003). This observation provides a fantastic opportunity: By examining how people create new concepts in a domain amenable to formalization in computational models, we can investigate (some aspects of) creativity, innovation, and imagination scientifically.

Towards the aim of providing a formal account of the cognitive mechanisms involved in creativity and innovation, we explore a simplified form of creative cognition that can be studied using standard behavioral methodology and is amenable to formal modeling: category generation. In a category generation task, participants are given some background knowledge about a domain (a cover story and/or learning one or more categories in the domain) and then are asked to generate another category in the domain. Unlike creativity and innovation, categorization is well studied from a breadth of perspectives, including behavioral, neural, comparative, and computational (Kurtz, 2015; Mack, Preston, & Love, 2013; Margolis & Laurence, 2015; Pothos & Wills, 2011). By analyzing a creative task using formal tools from the categorization literature, we can formalize some theories of creativity and test them empirically in a more rigorous manner.

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 (Regier, Kay, & Khetarpal, 2007; Rosch & Mervis, 1975; Shepard, Hovland, & Jenkins, 1961). Inspired by this work, previous research on the topic of category generation has explored a similar principle: People tend to create new categories that have similar *statistical regularities* as previously learned categories (Jern & Kemp, 2013; Ward, 1994). 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 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. Beyond the assumption that participant-generated categories are indeed novel, 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; Regier et al., 2007), the idea that a new category should be "different" is vague, as there are many ways it could be different from a previously observed category. Building on the largely successful exemplar modeling framework (Medin & Schaffer, 1978; Nosofsky, 1984, 1986), 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 approaches studied 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 classic series of reports, Ward and colleagues (Marsh, Ward, & Landau, 1999; Smith, Ward, & Schumacher, 1993; Ward, 1994, 1995; Ward, Patterson, Sifonis, Dodds, & Saunders, 2002) established that category generation is highly constrained by prior knowledge: Generated categories tend to consist of features observed in known categories, and they tend to exhibit the distributional properties as found in known categories. In a classic study, Ward (1994) asked participants to generate new species of alien animals by drawing and describing members of the species. People tended to generate species with the same features as on Earth (e.g., eyes, legs, wings), and possessing the same feature correlations as on Earth (e.g., feathers co-occur with wings). Likewise, aliens drawn from the same species tended to share more features with one another compared to members of opposite species.

NBC: More strongly link this to creativity: what do these findings indicate?

Much of the work from this area (e.g., Marsh et al., 1999; Smith et al., 1993) focuses on how sample cues (such as an example of a species generated by other participants) can drastically diminish creativity. However, 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. Theoretical accounts of these effects have primarily been grounded within the categorization literature. For example, the predominant "Path of Least Resistance" account (see Ward, 1994, 1995; Ward et al., 2002) proposes that, when generating a new species of animal, people retrieve from memory a known subcategory of animals (e.g., *bird*, *dog*, *horse*), and simply change some of the features to make something new. People are thought to change only features that are not characteristic of the retrieved category (e.g., if *bird* was retrieved, the presence of *wings* would not change, but *color* might). This theory incorporates elements of the highly influential basic-level categories framework (Rosch, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), as well as the exemplar view (Brooks, 1978; Medin & Schaffer, 1978). Problematically, however, while this work is been incredibly useful in providing a conceptual sketch of generation theories, the hand-drawn responses relied on in the experiments paradigms precludes the development of formal approaches.

Jern and Kemp (2013) recently showed that creative generation could be studied in a more controlled manner through the well-developed methods of an artificial categorization paradigm (see Kurtz, 2015, for a review). In Experiments 3 and 4 of their article, participants were exposed to members of experimenter-defined categories of "crystals" varying in size, hue, and saturation. Following a training phase during which the experimenter-defined categories were learned, participants were asked to generate novel categories of crystals. In a finding mirroring that of the Ward (1994) studies, Jern and Kemp 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 sampling model provided the strongest account. Their model views observed examples as samples from an underlying category distribution, describing the location of the category in the space, as well as how it varies along each feature. In turn, each category is viewed as a sample from an underlying *domain* distribution, specifying distributional commonalities among the observed categories. Generated categories are thought to stem from the same domain distribution as observed categories, thus the distributional properties of observed categories will be preserved within the generated category.

Jern and Kemp (2013) additionally tested a "copy-and-tweak" model that broadly resembles the earlier "Path of Least Resistance" account. The core proposal is that participants generate new items by copying stored examples from memory and tweaking them to generate something new. The copy-and-tweak model differs from the path of least resistance account in that it notably omits the hierarchical organization of categories, as well as selectivity in which features are changed. Instead, their copy-and-tweak model corresponds to a direct exemplar-similarity approach (e.g., Nosofsky, 1984, 1986): The model generates new items according to their similarity to known members of the target category. This model provided a poor account of the observed data, as the experiments devised by Jern and Kemp were specifically designed to challenge this model. However, its application is notable as a first step toward explaining category generation using well-known formal approaches from the categorization literature.

2.1 Something Different: A Role For Contrast

From the perspective of the categorization literature, 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 towards 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, contrast should be viewed a primary constraint on creative generation: new concepts must 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 sampling model developed by Jern and Kemp (2013) assumes that differences between generated category 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 authors do not make predictions about the *location* of the category within the domain (the perceptual instantiation of category members). Indeed, under a strict interpretation of their model, the most probable new category to be generated is located in *exactly* the same location and distributional information as the given category. However, this is not simply an issue with their model but with the broader class of standard hierarchical Bayesian models. 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 is due to *noise* and should not be *systematic*. The best a hierarchical Bayesian model can do at capturing contrast is to assume that the new category is placed uniformly at random over stimulus space. This defeats the purpose of a hierarchy as new category locations do not depend on the locations of previously observed categories.

The copy-and-tweak model tested by Jern and Kemp (2013) also claims little about how generated categories should contrast with what is already known. In their simulations, the model was only tested on generation after the learner had been exposed to members of the target category, and so the model's ability to generate a new category from scratch was

not evaluated. However, the model's generation is based exclusively on similarity to known members of the target category; when there are no members of the target category, generation is presumably random.

2.1.1 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 is 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 in creative generation, we conducted a novel analysis of Experiment 3 from Jern and Kemp (2013).

NBC: Does it feel strange to focus on Experiment 3 specifically? Do you think wee need to explain why this is the 'right' experiment for our question?

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, Jern and Kemp (2013) 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. Jern and Kemp (2013), however, did not analyze or discuss how the generated categories *differed* from the experimenter-defined categories.

Because each experimenter-defined category in the Jern and Kemp (2013)

experiment possessed a distinct hue shared by all members of the category, it is sensible to propose that participants would 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. Problematically, however, stimulus hue was presented using 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; Wyszecki & Stiles, 1967). Further, they did not calibrate their monitor, and so we cannot know the precise values of colors presented to participants. As Jern and Kemp (2013) were interested in relations between the saturation and length of examples in generated into novel categories, this is not an issue for 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.

Although we cannot know precisely which colors 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 Jern and Kemp (2013) 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. By categorizing the participant-generated crystals likewise, we can obtain a broad measure of category contrast by determining the proportion of participant-generated crystals that fall into the same groups as the experimenter-defined categories: If contrast influences the hues of the

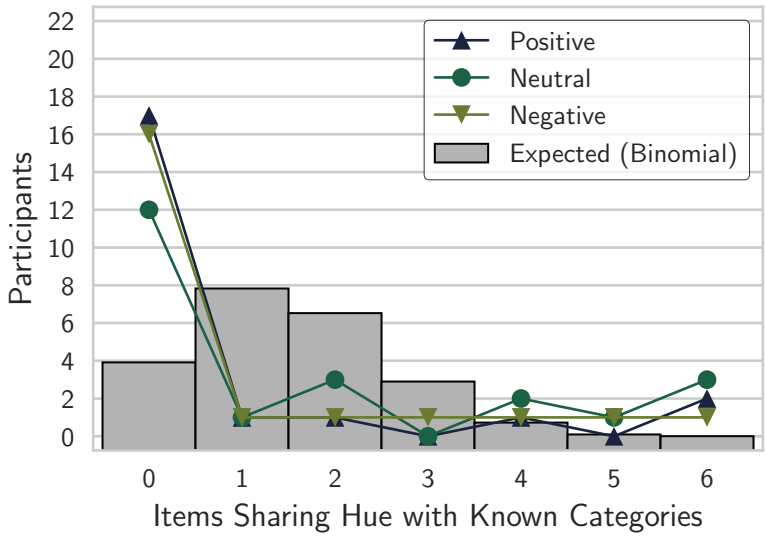

Figure 1: Analysis of data from Jern and Kemp (2013), Experiment 3. Plotted is the number of generated items that share a color group with one of the experimenter-defined classes. The "Expected" data follows a Binomial distribution with $p = 2/8$, given there were two experimenter-defined classes, and eight color groups.

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 $\tilde{\chi}^2 > 200$, $p < 0.001$): 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 of Jern and Kemp (2013) provides preliminary support that contrast plays a role in category generation. Taken alongside the analyses reported by Jern

and Kemp (2013), 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 somewhat arbitrary, and it is not clear that our color grouping is consistent with psychological color boundaries. While we did obtain similar results using a variety of alternative groupings, the hue dimension used in the Jern and Kemp (2013) 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. 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.

2.2 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: generated categories 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 category learning (GCM; Nosofsky, 1984, 1986).

Both PACKER and the GCM simulate categorization under the assumption that learners represent categories as a collection of exemplars, corresponding to the labeled

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

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

where w_k is the attention weighting of dimension k ($w_k \geq 0$ and $\sum_k w_k = 1$), accounting for the relative importance of each dimension in similarity calculations, and c ($c > 0$) is a specificity parameter controlling the spread of exemplar generalization. For simplicity, 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 (see Garner, 1974; Shepard, 1964). In our simulations, we set $r = 1$ due to the separable nature of the stimulus dimensions used in our experiments (see Figure 3).

PACKER (as well as its name) was in part inspired by earlier work from the categorization literature (Hidaka & Smith, 2011). They argued that natural categories "pack" the values of features such that different categories fill the space with distance between one another, while maintaining 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 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.

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

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

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

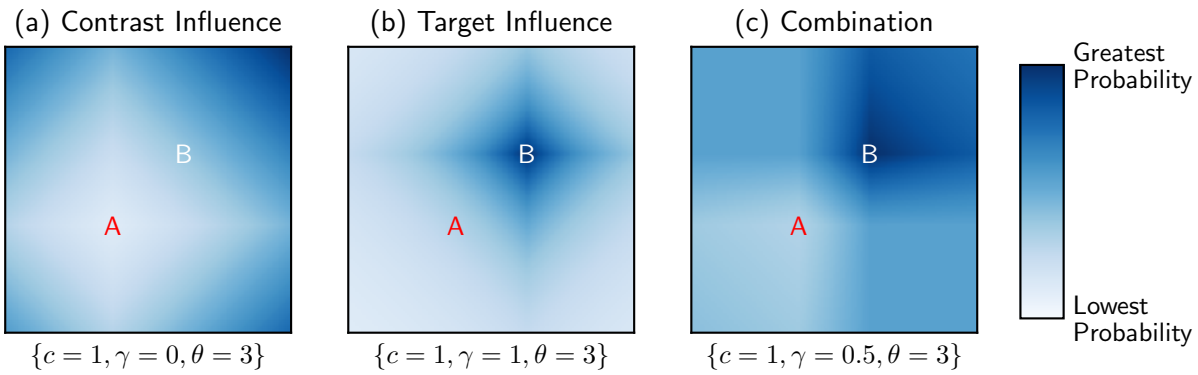

Figure 2: PACKER generation of a category 'B' example, following exposure to one member of category 'A' and one member of category 'B'. Predictions are shown for three different parameterizations (differing only in γ): *(a)* Predictions based on contrast similarity only. *(b)* Predictions based on target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

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 Jern and Kemp (2013) when $\gamma = 1$. Likewise, when $\gamma = 0$, PACKER represents 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, although it may be possible to extend Jern and Kemp (2013)'s hierarchical sampling model to include contrast, this mechanism would be at odds with a 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 aggregated similarity slightly differently. 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.

2.2.1 Relation Between PACKER and Copy-And-Tweak

The PACKER model is similar to the copy-and-tweak model reported by Jern and Kemp (2013): 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 Jern and Kemp (2013). 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 Jern & Kemp, 2013) for future work.

2.3 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 provided encouraging initial support that contrast influences category generation (Jern & Kemp, 2013, 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 Jern and Kemp (2013)'s paradigm: 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 sampling and copy-and-tweak models developed by Jern and Kemp (2013).

3 Experiment 1

To begin our investigation, we sought to extend the early evidence obtained from our analysis of the Jern and Kemp (2013) data, under a variety of learning conditions and using more standard stimulus materials. We used an artificial stimulus design, 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., Conaway & Kurtz, 2016a, 2016b; Nosofsky, Gluck, Palmeri, & McKinley, 1994; Shepard et al., 1961).

Unlike those used in the Jern and Kemp (2013) experiments, proximity along these 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 Jern and Kemp (2013) 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.

NBC: can you think of other studies with size and color of squares as features?

Panels B-D of Figure 3 show the locations of exemplars belonging to the experimenter-defined categories ('A', or 'Alpha') that participants were assigned to learn about prior to generating a new category. 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 variation along each dimension (e.g., $dark \rightarrow light$ or $light \rightarrow dark$) 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

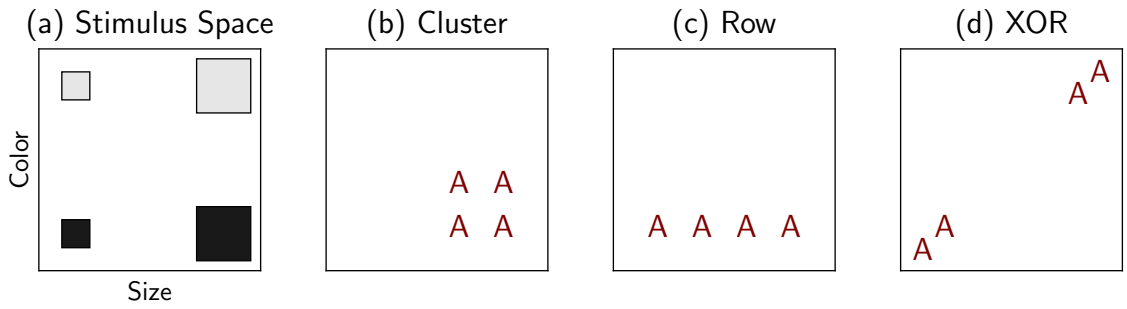

Figure 3: Stimulus domain and category types tested in Experiment 1. Stimuli are not drawn to scale.

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).

In this way, we can analyze the results of the first experiment as a conceptual replication of the classic finding that generated categories tend to share distributional properties with known categories in the domain (see Jern & Kemp, 2013; Ward, 1994). From these results, we can predict that, in each condition, participants should generate categories that are distributionally similar to the experimenter-defined category: In the Cluster condition, generated categories should be tightly clustered. In the Row condition, generated categories should vary more along the X-axis than the Y-axis. In XOR condition, generated categories should be widely distributed across both dimensions, and the two dimensions should be positively correlated.

Interestingly, the XOR condition also offers a dissociation between the roles of category contrast and the emulation of distributional structure: widely-distributed,

positively-correlated categories would need to lie along the positive diagonal of the space (that is the only place they "fit"), which is already occupied by the experimenter-defined category. Thus, if contrast plays a role, exemplars in the generated categories of participants in the XOR condition may not be positively correlated – they may be negatively correlated instead.

3.1 Participants & Materials

183 participants were recruited from Amazon Mechanical Turk. Participants were 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. Stimuli were squares varying in color (RGB 25–230) 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.

JLA: we should probably mention the discrepancy between the number of participants per condition and explain why it isn't anything to worry about.

3.2 Procedure

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

Following the training phase, participants were asked to generate four examples

belonging to another category called 'Beta'. As in Jern and Kemp (2013), generation was completed using a sliding-scale interface. Two scales controlled the features (color, size) of 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.

3.3 Results

To begin, it is worth noting that 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 evident that different participants generated qualitatively different category structures. In this section we will focus on what can be learned from 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

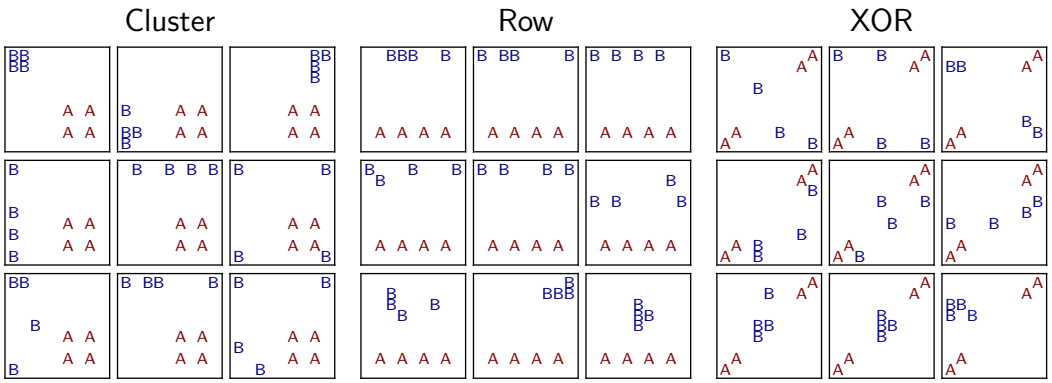

Figure 4: Sample categories generated by participants in Experiment 1. Representative samples from common generation profiles are shown.

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.

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: all possible generation candidates are distant from at least one cluster of the experimenter-defined category. These data reveal a systematic pattern: the majority of participants generated categories with more 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 formally evaluate this claim, 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 clear evidence of an effect of category contrast: participants prefer to generate categories

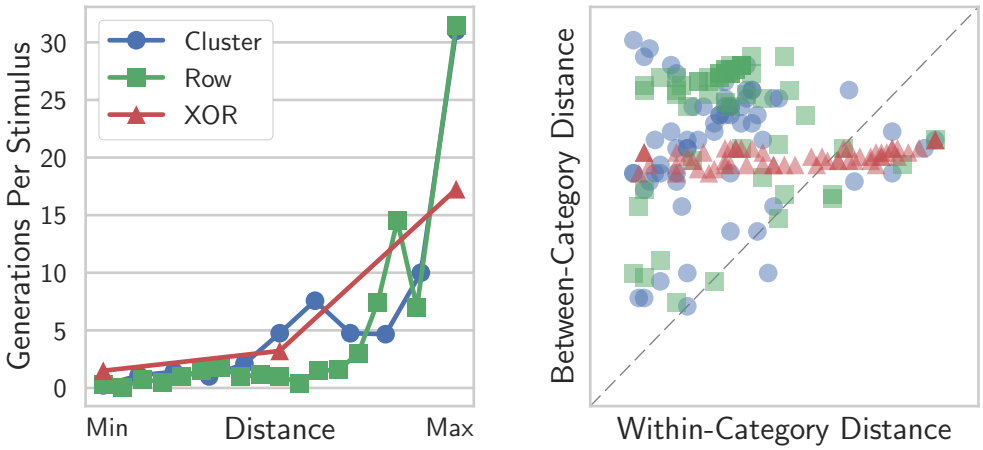

Figure 5: Experiment 1 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

that are dissimilar to the learned category but maintain some level of internal cohesion.

A secondary goal of this experiment was to examine whether we replicate the classic result that generated categories often possess the same distributional properties as previously-known categories. 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-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,

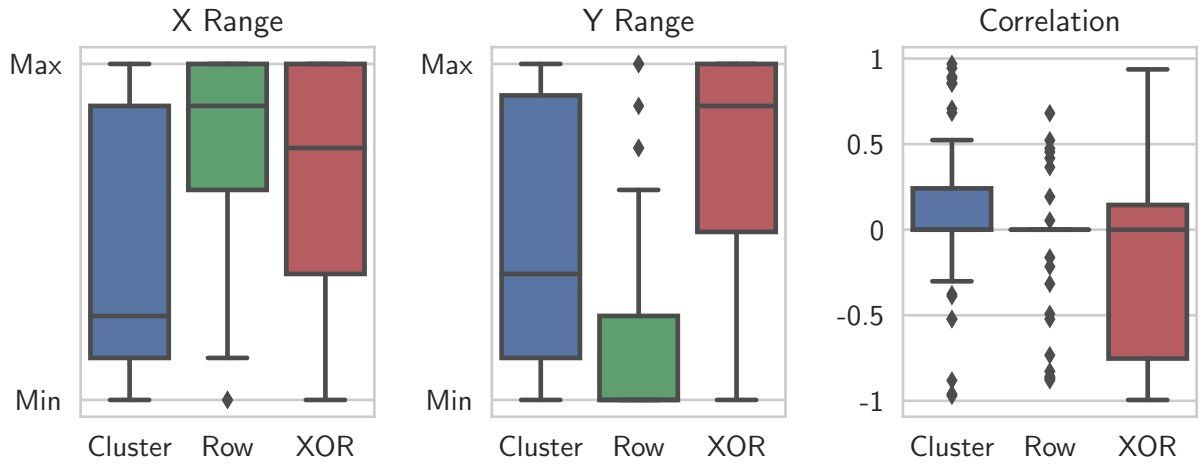

Figure 6: Box-plots of the distributional statistics from the categories generated in Experiment 1. Boxes depict the median and quartiles of each condition, with whiskers placed at 1.5 IQR. All points outside this region are marked individually.

$t(117) = 2.56$, $p = 0.012$. This latter effect was not expected; 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). This indicates that there is more to category generation than the distributional structure of other categories in the domain. Further, as we will discuss in more detail in the model-based analysis section, it is expected by our proposal that contrast is a fundamental principle in category generation.

3.4 Discussion

In Experiment 1 we sought to extend our analysis of the Jern and Kemp (2013) 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 (Jern & Kemp, 2013; Ward, 1994): members of generated categories were more widely distributed along dimensions which were widely distributed in the experimenter-defined category.

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

Thus, while the core results of Experiment 1 indicate that generated categories systematically contrast with what is already known, the negative correlations observed in the XOR condition may signal something further. That is, the constraints on creative

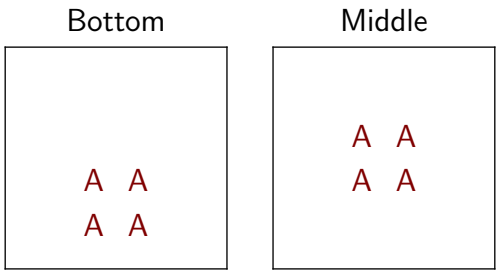

Figure 7: Category types tested in Experiment 2.

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.

4 Experiment 2

In order 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 an identical distributional structure (both are tight clusters of examples with a neutral feature correlation), and only differ slightly 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 structure of previously learned categories were the only influence on the structure of generated categories, we should observe no difference between these two conditions with respect to distributional structure.

Critically, however, participants seeking to make a perceptually distinct category would be more likely to distribute members of the generated category into areas that are distant from members of known categories. Thus, 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 types produces a considerable change to the shape of the unoccupied space: participants assigned to learn the Bottom category would 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, to the extent that category contrast influences distributional structure, we should observe more participants in the Middle condition that generate examples above *and* below the experimenter-defined category.

4.1 Participants, Materials, & Procedure

122 participants were recruited from Amazon Mechanical Turk. 61 participants were randomly assigned to the Middle and Bottom conditions each. The stimuli and procedure were exactly as in Experiment 1: participants first completed a short training phase, followed by the generation phase.

4.2 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 is organized into four columns based on commonly observed patterns of generation: a 'Cluster' type of tightly-clustered examples, 'Row' and 'Column' types of exemplars widely distributed along the one axis but narrowly along the other, and a 'Corners' type, wherein participants placed exemplars in disparate corners of the space. As before, in this section we focus on

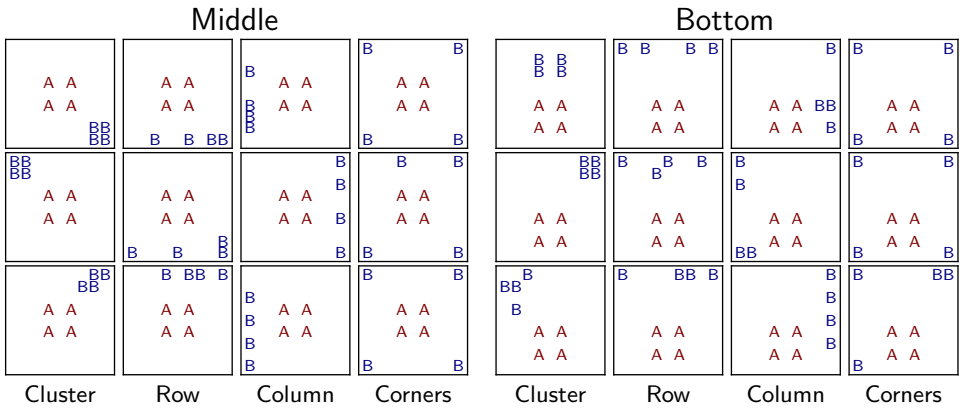

Figure 8: Sample categories generated in Experiment 2.

what can be concluded from analyzing the data in aggregate, but in later sections we will focus more specifically on explaining the individual differences.

We began our analysis by again 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 highly similar results. Stimuli that are more distant from members of the experimenter-defined category were more frequently generated, 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, $t(60) = 2.71$, $p = 0.009$.

We did, however, observe a notable subgroup of participants in each condition who generated categories with more within-category than between-category distance. Upon manual inspection, many of these individuals appear to have assumed a 'Corners' strategy, placing exemplars in disparate corners of the space, thus producing much more within-category distance, see Figure 8 for examples.

To address out main question, we followed this analysis by comparing the conditions on measures of distributional structure. For each generated category, we computed the range of exemplars along each axis (X, Y), as well as the correlation between features.

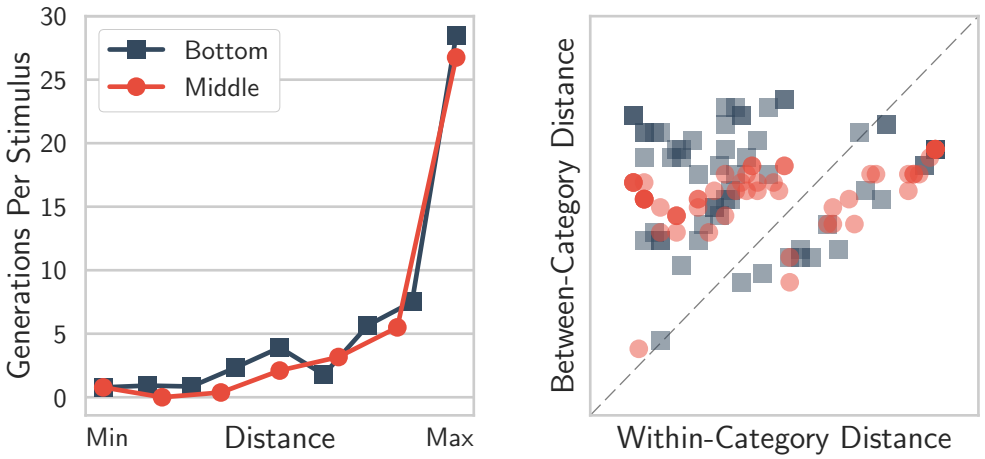

Figure 9: Experiment 2 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-category versus between-category distance in each of the participant-generated categories.

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 in turn be less likely to generate category members in the bottom regions because members of the experimenter-defined category are located there. These predictions can be addressed within the Y-axis ranges of the generated categories: If Middle participants utilize the upper and lower regions of the space, their categories should vary more along the Y-axis. T-Tests comparing the conditions on the distributional statistics, however, reveal few between-group differences: the conditions do not differ with respect to X-axis range, Y-axis range, or feature correlations ($ps > 0.17$).

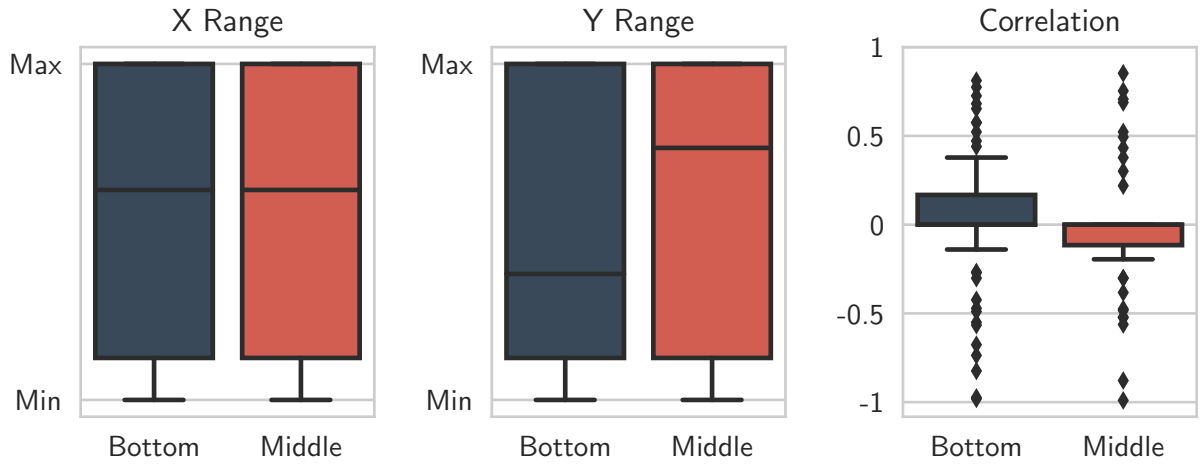

Figure 10: Box-plots of the distributional statistics from the categories generated in Experiment 2.

However, our ability to detect differences between the conditions is, in this case, diminished due to the non-normality of the data. 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.

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

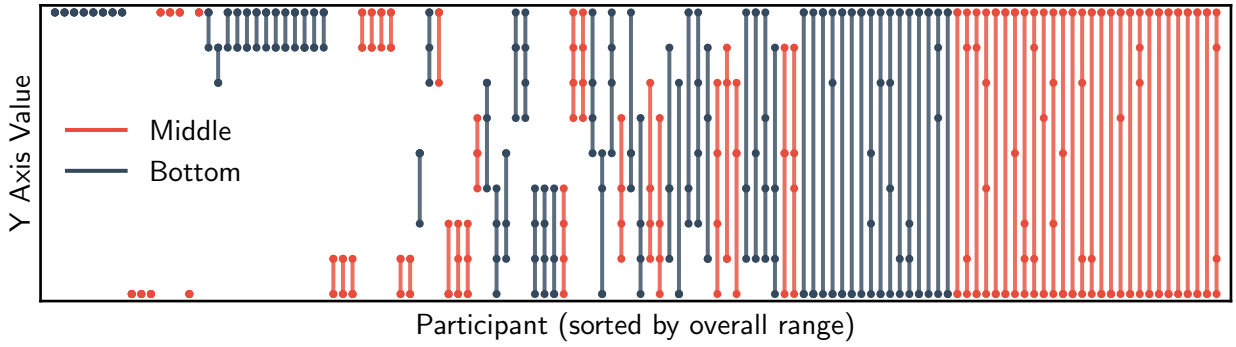

Figure 11: Y-Axis range and position of the categories generated in Experiment 2. Each line corresponds to a single category, with notches corresponding to the Y-axis position of exemplars within the category (notches may overlap). Data from each participant is sorted by overall range, and then by condition.

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

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.

4.3 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 under the viewpoint that the distributional structure of previously learned categories is the sole determinant of the distributional structure of generated categories. However, the observed performance is sensible given 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.

5 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 2.2.1), 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 sampling model proposed by Jern and Kemp (2013), described in-depth in Appendix A. The comparison between the hierarchical sampling model and PACKER is meant to emphasize the necessity of contrast and demonstrate it cannot be explained by emulating of distributional structure: whereas PACKER is insensitive to the distributional structure of learned categories (relying only on within- and between-category similarity), the hierarchical sampling 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 we more specifically describe the strengths and weakness of each model's account.

5.1 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

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

PACKER	**Copy & Tweak**	**Hierarchical Sampling**
$AIC = 9095$	$AIC = 9842$	$AIC = 9912$
$L = -4545$	$L = -4919$	$L = -4952$
$c = 0.482$	$c = 3.187$	$\kappa < 0.001$
$\gamma = 0.525$	$\gamma = 1$ (fixed)	$\nu = 5.596$
$\theta = 6.664$	$\theta = 2.969$	$\lambda = 0.055$
		$\theta = 3.174$

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 2.2), as well as four in the hierarchical sampling 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; Akaike, 1974), where smaller values correspond to better fits (discounted by the number of parameters). 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 sampling models, and copy-and-tweak performed somewhat better than the hierarchical sampling 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 sampling model may be attributed to its

exemplar-based representation, as opposed to the prototype-based representation assumed in the hierarchical sampling 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 used to represent categories in the Jern and Kemp (2013) model, but can be handled easily using an exemplar-based approach.

A key distinction between PACKER and copy-and-tweak, as well as the hierarchical sampling 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 mean of the category distribution, the hierarchical sampling 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). 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 sampling ($L = -3612$, $AIC = 7232$) 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 ($\chi^2(1) = 748, p < 0.001$ for all data and $\chi^2(1) = 378, p < 0.001$ excluding the first example), providing more targeted evidence that category generation is better explained when contrast is considered.

NBC: report something about robustness to model comparison metric (BIC and AIC_c).

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. Plotted is PACKER's log-likelihood as a function of γ parameter.

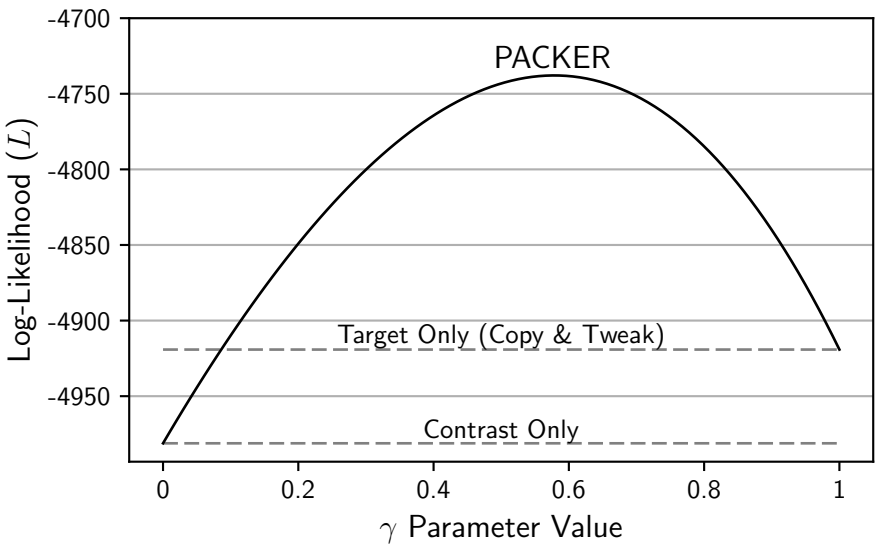

Figure 12: PACKER's fit as a function of its prioritization of within-category and between-category similarity (using the γ parameter). To facilitate comparison, PACKER's other parameters (c, θ) were set to the best fitting values obtained for copy-and-tweak in Table 2.

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 other parameters are fitted, as in Table 2). Notably, however, the copy-and-tweak parameterization ($\gamma = 1$) performs better than the 'contrast-only' parameterization of the model ($\gamma = 0$). Thus, it is evident that the data is better explained when both within-category similarity and category contrast is considered.

5.2 Individual Differences

As noted in Experiments 1 and 2, we observed a great deal of individual differences in the types of categories generated. Within each condition, we observed tightly clustered and widely distributed categories, as well as row and column categories, and so forth (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 targeted within the PACKER framework, simply by tuning the model's parameters. In this section, we describe more specifically how the most frequently observed profiles can be realized.

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

In the framework proposed by PACKER, the cluster and corners profiles arise based on different prioritization of within-category similarity versus between-category contrast, and the row and column profiles arise based on the prioritization of each dimension in the computation of similarity. For example, in the cluster profile, there is a high degree of within-category similarity along both dimensions, whereas in the corners profile there is minimal within-category similarity. Thus, PACKER's proposal is that these individual differences arise as a result of different priorities: While the tight cluster configuration can be considered PACKER's 'default' mode (as it maximizes within-category similarity), the corners profile can be produced when between-category contrast is put at a higher priority (i.e., $\gamma = 0$).

Likewise, in the row and column profiles, there is a high 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 (see Equation 1). Traditionally, the attention

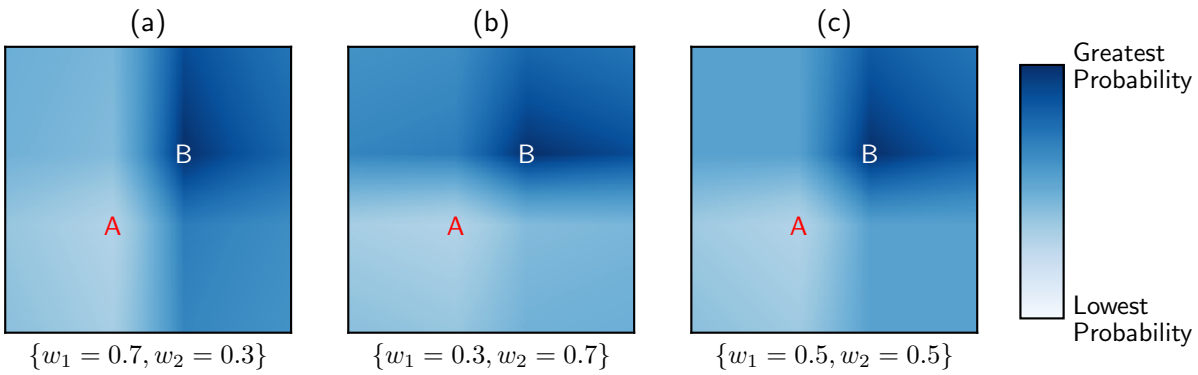

Figure 13: PACKER generation of a category 'B' example, following exposure to one member of category 'A' and one member of category 'B'. Predictions are shown for different attention settings: *(a)* Increased weighting of the X-axis. *(b)* Increased weighting of the Y-axis. *(c)* Uniform weighting (identical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

weights in exemplar models are thought to reflect the diagnostic value of each dimension towards classifying the known category members (Kruschke, 1992; Nosofsky, 1984, 1986), but within a generation context the weights specify the importance of within- and between-category similarity along each dimension. For example, if 100% of attention is allocated along the X-axis, 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 counted 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.

Like in PACKER, changes to the weighting of the dimensions can also be used to produce row- and column-like categories in the copy-and-tweak and hierarchical sampling models. Indeed, as copy-and-tweak is simply a special case of the PACKER model, the attention weights operate exactly as described above. Within the hierarchical sampling model, the covariance matrix specifying the prior domain distribution, Σ_0, can be used to similar effect. 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.

However, a key limitation within both in both of these models is the lack of interaction between the distributional structure of a category (in this case, a row-versus-column orientation) and the location of that category within the domain. While these models are *capable* of generating row and column categories, there is no mechanism in place to ensure generated categories will be distinct from what is already known. In the section below, we explore the interdependence between distributional structure and location in creative generation.

5.3 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 sampling models do not make any claims about how generated categories should differ from what is already known. However, as we observed in Experiment 2, the distributional structure of a category is not always independent of its location within the domain. To demonstrate this point in a broader manner, 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.

For each possible stimulus, we compiled the range differences across all the categories it was generated into. However, because many stimuli were infrequently generated (such items near members of the 'Alpha' category), we cannot simply compute the empirical average, as infrequently generated stimuli would be likely to show artificially strong differences. Instead, we aggregated the scores in an empirical Bayesian manner (Robbins, 1964): the aggregated score for each stimulus was viewed as a normal distribution with a $\mu_0 = 0$ and σ set based on the observed scores x for that stimulus. The aggregated difference μ_x was then computed as a Bayesian update:

$$\mu_x = \frac{\mu_0 + \sum x}{1 + n_x} \tag{4}$$

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 both strongly positive and negative scores, are given neutral difference scores. The results of our analysis[1] are shown in Figure 14.

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.

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

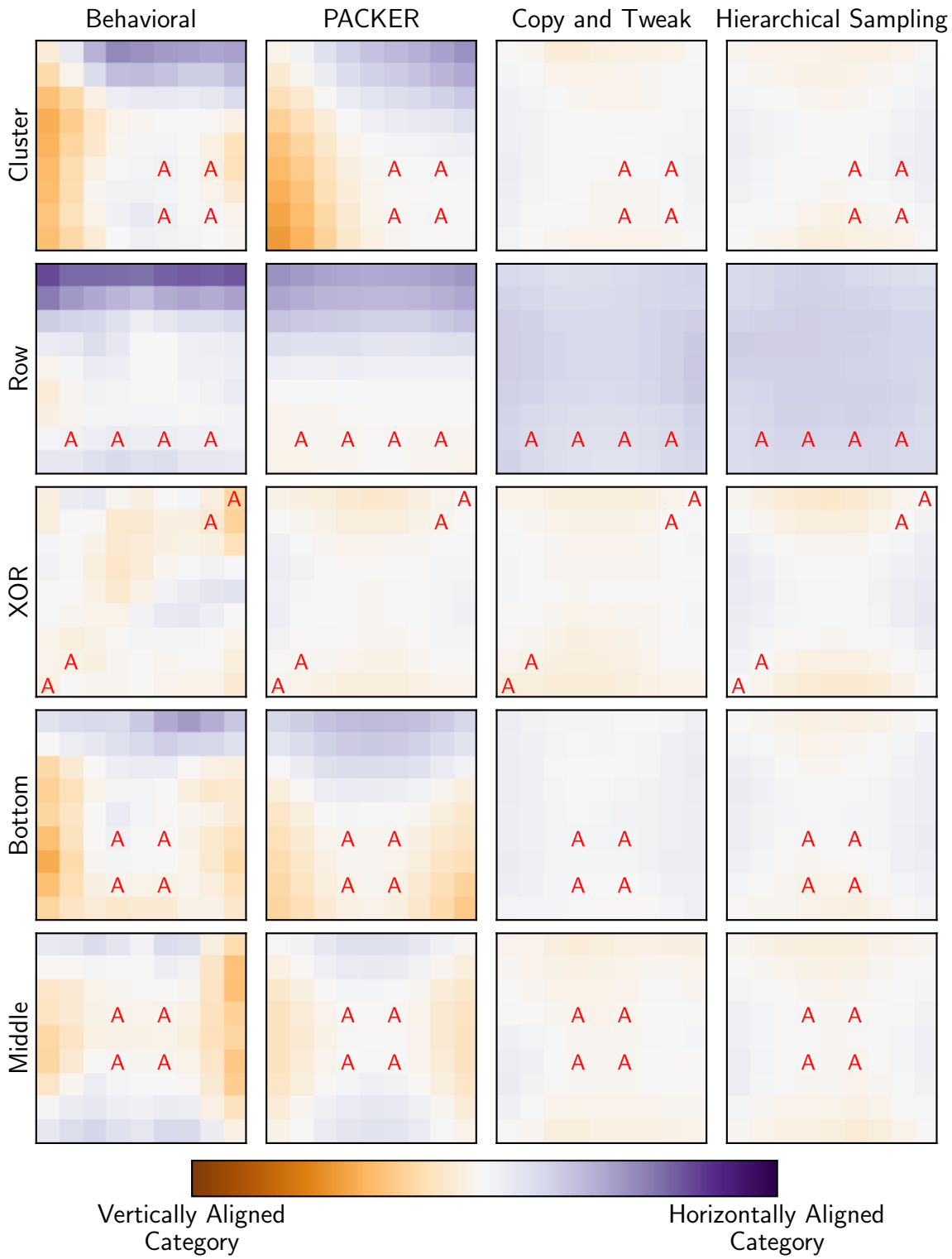

Figure 14: Behavioral and simulated range difference gradients. Each panel shows, for each stimulus, the dimensional orientation of the categories it was generated into: vertically aligned 'columns' versus horizontally aligned 'rows'.

NBC: Need to run this on lando with many more samples. current = 30

JLA: This is beautiful. I'd like to add one thing to it. On the left margin have something that says "Experiment 1" and groups the top 3 rows, and then "Experiment 2" for the bottom two.

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

To attempt to explain our findings within the PACKER, copy-and-tweak- and hierarchical sampling models, we conducted simulations using an individual-differences approach. As noted in Section 5.2, row- and column-like categories can be produced by within model through changes to the weighting of each dimension. Given this information, we may use the models to simulate each participant's generation separately, with the importance of each dimension set according to the relative range of the participant's generated category along each dimension.

In the PACKER and copy-and-tweak models, the attention weights, w, specify the importance of each dimension in the computation of similarity. While there exist methods to find the optimal attention weighting scheme given a classification (see Vanpaemel & Lee, 2012), for simplicity we 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:

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

where θ_w is a free parameter controlling how differences in range correspond to differences in weights (functioning similarity to the θ parameter in each of the models), and range(k) is the range of examples generated by the participant along dimension k. In our simulations $\theta_w = 1.5$, 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.

In the hierarchical sampling model, the covariance matrix specifying the prior domain distribution, Σ_0, can be used to similar effect. 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.

Unlike the PACKER and copy-and-tweak models, the hierarchical sampling 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 sampling model, we computed the dimensional variances according to: $\lambda(1 - w_k)d$, where λ is a free parameter specifying the overall assumed variance of the domain, and d is the number of dimensions. 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.

NBC: This works for 2D, but not 3D and beyond (as $1 - w_k$ would not produce a valid weighting).

JLA: I renamed λ because λ is reserved for the correlation parameter of a covariance matrix (especially a 2×2).

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 **NUMBER** of times, resulting in **NUMBER * PARTICIPANTS** categories generated by each model. 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 procedure described above, PACKER provided a much closer match to our behavioral results than the copy-and-tweak and hierarchical sampling 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 sampling 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, in which the majority of participants generated a 'Row'-like category, widely distributed along the X-axis but not the Y-axis. In these case, both models are initialized with weights that produce Row categories, but because category contrast is not considered, categories are uniformly generated across the entire domain, rather than concentrated within the upper-regions as observed behaviorally.

6 General Discussion

The creative generation of novel concepts is an intriguing yet understudied topic in cognitive science. 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? This issue is of fundamental importance: in order to successfully generate something new, it must be different from what is 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 (Jern & Kemp, 2013, 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. Finally, we conducted simulations comparing PACKER's account of our results to that of a "copy-and-tweak" model (realized as a variant of PACKER with no sensitivity to category contrast), and a hierarchical sampling model designed to explain the classic distributional similarity effect. In all simulations, we found that PACKER's sensitivity to contrast considerably enhanced the models account of generation.

The PACKER model bridges two gaps in the literature. First, it provides a quantitative theory of the role of contrast in creative generation, a topic more or less overlooked by previous work. The lack of attention to this issue is likely due to the underconstrained notion of what it means for two concepts to be "different". In PACKER, this notion is interpreted using the theory of similarity employed by the exemplar view of categorization (Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1984, 1986). This theory of similarity, tracing its foundations to classic research in stimulus generalization (Attneave, 1950; Shepard, 1957, 1987), has had broad influence in many fields within cognitive science. Thus, beyond providing a formal specification of category contrast, PACKER also aligns the relatively understudied topic of creative generation more closely with the well-developed literature on human categorization. The success of the PACKER model indicates that core theories from the categorization literature may offer a strong account of how people generate new concepts.

The success of PACKER yields reveals keys insights about the nature of creative generation. First, 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: it is clear that both constraints influence creative generation.

This finding is extended by our observation of the broad interdependence between the distributional structure (feature variance, correlation) and physical instantiation (location within the domain space) of generated categories. In Experiment 2, we found that the unoccupied regions of the domain influenced the distributional structure of categories, and in Experiments 1 and 2 we observed that generated categories assume different distributional structures depending on how they contrast with known categories. This interdependence is straightforwardly explained under PACKER's dual-constraint proposal: after first establishing a location for their generated category, participants distribute exemplars in order to minimize between-category similarity while not drastically decreasing within-category similarity.

However, PACKER does not provide a full account of what is known about category generation. Most notably, in this paper we have not evaluated the model's ability to explain the classic finding that generated categories tend to share distributional commonalities with previously learned categories (see Jern & Kemp, 2013; Marsh et al., 1999; Smith et al., 1993; Ward, 1994, 1995; Ward et al., 2002). 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 sampling model indicates that the emulation of distributional structure exerts only a limited influence in comparison to category contrast.

Nonetheless, these classic effects are a core element of the phenomenology of creative generation, and it is worth noting that PACKER does not include any mechanisms explain them. Instead, through the development and evaluation of the PACKER model, we have sought to add new elements into such a phenomenology: specifically, 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 sampling approach proposed by Jern and Kemp (2013) with PACKER's underlying claims to obtain a "best of both worlds" model, capable of explaining the role of contrast in 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 the hierarchical Bayesian approach. Thus, while such a model may provide an excellent fit to the data, its theoretical composition would be more or less of an amalgam, and thus it may not succeed in other areas of model evaluation (such as predicting and explaining novel phenomena).

As it stands, however, PACKER's mathematical basis is subject to a wide variety of interpretations, leading to distinct views on the nature of creative generation.

JLA: This is pretty disorganized at the moment. There's a lot of good in here though. Here's the structure I'd like to see. Please move and reorganize it into these subsections

Summary of results

6.1 Implications for creative cognition

more on the general theories/philosophy of creative cognition and what we bring to the table.

6.2 Contrast in categorization

More broad perspective of how this fits with categorization more broadly. Similar style model for explaining language/perceptual (phonemes and color) categories. Also, you should check on studies of how contrast affects categorization more broadly (i think there's some work on that).

6.3 Limitations and Future Directions

this is where the importance sampling connection will come in and will involve specifying the computational problem precisely. that hasn't been done yet for category generation and then perhaps to creativity more broadly.

7 Conclusions

8 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.

References

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

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

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

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

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

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

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual Organization in Vision: Behavioral and Neural Perspectives* (p. 233-278). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

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

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

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*, 2023-2027.

Margolis, E., & Laurence, S. (2015). *The conceptual mind: New directions in the study of concepts*. Cambridge, MA: MIT Press.

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

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

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

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

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

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

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

Robbins, H. (1964). The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, *35*(1), 1–20.

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

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

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

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

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

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

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

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

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

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

Ward, T. B. (1995). What's old about new ideas. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 157–178). Cambridge, MA: MIT Press.

Ward, T. B., Patterson, M. J., Sifonis, C. M., Dodds, R. A., & Saunders, K. N. (2002). The role of graded category structure in imaginative thought. *Memory & Cognition*, *30*(2), 199–216.

Wyszecki, G., & Stiles, W. S. (1967). *Color science*. New York: John Wiley.

A The Hierarchical Sampling Model

Jern and Kemp (2013) demonstrated how a hierarchical Bayesian model could explain the distributional correspondences between observed and generated categories. In their model, exemplars of generated category 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, each category's exemplars are assumed to be sampled from a multivariate Normal distribution with parameters (μ, Σ). Each category's covariance matrix is assumed to be inverse-Wishart distributed with parameters (v, κ, and Σ_D).[2] Σ_D is the covariance matrix shared between categories. We assume the shared covariance matrix Σ_D is generated from a Wishart distribution (for conjugacy) with parameters v_0, κ_0, and Σ_0. We set $\nu_0 = 4$, and $\Sigma_0 = \lambda\mathbf{I}$, where λ is a free parameter controlling the expected variance of dimensions (dimensions of the shared covariance matrix are expected to be uncorrelated) and $\mathbf{I}$ is the identity matrix.

NBC: check on that $\nu_0 = 4$ business

To simplify the model predictions, we used *maximum a posteriori* (MAP) estimates for the hidden parameters and then generated new categories based on those estimates. Due to conjugacy, the MAP estimate for the shared covariance matrix $\Sigma_D = \Sigma_0 + \sum_c C_c$, where C_c is the empirical covariance matrix of category c. The MAP estimate of the covariance

[2]Note that Jern and Kemp (2013)'s model is slightly different, as they used a non-conjugate model. Their model acts very similar to our version of it and receives comparable fits.

matrix for the target category B is

$$\Sigma_B = \left[\Sigma_D \nu + C_B + \frac{\kappa n_B}{\kappa + n_B}(\bar{x}_B - \mu_B)(\bar{x}_B - \mu_B)^T\right](\nu + n_B)^{-1} \tag{A.1}$$

where ν ($\nu > k - 1$) is an additional free parameter (from the Inverse-Wishart prior on Σ_B) weighting the importance of Σ_D. When the target category has no members (i.e., $n_B = 0$), items are generated at random.

NBC: should explain Equation A.1 more fully now that we are not limited for space.

JLA: K, though if I remember, it's just conjugacy, in which case we can just cite a paper that has IW-W conjugacy.

NBC: It **is** just conjugacy, so we can just cite a paper (do you know of a paper or should I start digging?). We'll probably want to detail the what we did for priors and which free parameters we fitted though.

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

$$p(y) = \frac{\exp\{\theta \cdot \text{Normal}(y; \mu_B, \Sigma_B)\}}{\sum_i \exp\{\theta \cdot \text{Normal}(y_i; \mu_B, \Sigma_B)\}} \tag{A.2}$$

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