Creating Something Different: Similarity and Contrast in Concept Generation.

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

[1]University of Wisconsin-Madison, Department of Psychology, 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

1 Abstract

JLA: Joe! Did you know i made a command for you to enter in comments like this??

JLA: This is awesome!!!

```
\jlanote{Some note}
```

Keywords: categorization, concepts, creativity, generation

2 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, scientific breakthroughs are rare and the result of a great deal of cognitive effort, which it difficult to investigate the underlying cognitive representations and processes responsible for them using standard experimental methodology or 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 a novel sentence, thought, and/or drawing.[1] 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 (Jern & Kemp, 2013; Ward, 1994). Unlike creativity and innovation, categorization is well studied from a breadth of perspectives, including behavioral, neural, comparative, and computational perspectives (Kurtz, 2015; Mack, Preston, & Love, 2013; Margolis & Laurence, 2015; Pothos & Wills, 2011). By analyzing a creative task using formal tools from the

[1] **JLA**: We need citations/examples for this. In a compositional system, it's pretty clear this has to be the case, but I'd like to use someone else as a reference rather than arguing this myself.

categorization literature,

Previous work has focused on one key aspect of category generation: People tend to create new categories that have similar *statistical regularities* to previously known 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 be one that is identical to a prototypical previously observed category (and thus, not new at all).

In this article, we propose a new constraint that guides category generation: "being different" or contrasting from other categories in the relevant domain. To some extent, this constraint has been implicitly assumed in much previous work: 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.

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.

3 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 & 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.

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 & 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 were 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, the experimental paradigms employed in these studies were relatively uncontrolled, precluding the development of formal approaches. Specifically, participants in these studies are typically asked to generate objects that they have a rich degree of prior knowledge about (e.g., plants, animals, tools, toys), and so it is difficult to apply formal models to the data.

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). In their experiments 3 and 4, 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 Bayesian hierarchical 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): participants are thought to generate new items according to their similarity to known members of the target category. This model provided a poor account of the observed data: indeed, 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.

3.1 Something Different: A Role For Contrast

It is worth noting that the literature reviewed above provides a somewhat limited view of 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 modeling efforts have been towards explaining those effects. In this paper, we investigate another important constraint on the creative use of conceptual knowledge: category contrast. In order to generate a novel concept, individuals must produce something that is in some capacity *different* from what they already know. Thus, in a trivial sense, contrast can be viewed as a fundamental constraint on creative generation: new concepts must firstly 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), for example, does not provide any strong claims about how generated categories should contrast with learned ones: the model only proposes 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 the hierarchical Bayesian approach presented by Jern and Kemp, the most probable new category to be generated is located in *exactly* the same location as the given category. However, this is not simply an issue with their model – Hierarchical Bayesian models assume 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 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 was the case in the simulations conducted by Jern and Kemp).

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 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.

3.2 The PACKER Model

As noted above, the constraint that new concepts should differ from what is already known has been largely been 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 formalizing the role of contrast in creative generation, we developed the PACKER (*Producing Alike*

and Contrasting Knowledge using Exemplar Representations) model which 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).

NBC: Maybe note something about how PACKER's basis in the exemplar view means that we're explaining category generation using well known ideas, and that exemplar models are viewed as highly principled.

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 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 2).

PACKER's core proposal (as well as its name) was in part inspired by earlier work from category learning literature (see Hidaka & Smith, 2011). 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 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. PACKER sets $f(x_j)$ depending on exemplar x_j's category membership: $f(x_j) = \phi$ if x_j is a member of a contrast category, and $f(x_j) = \gamma$ if x_j is a member of the target category. ϕ and γ are free parameters ($-\infty \leq \phi, \gamma \leq \infty$) controlling the contribution of contrast- and target-category similarity, respectively. Larger absolute values result in greater consideration of those exemplars, with values of 0 eliminating their effect. A negative value for $f(x_j)$ produces a 'repelling' effect (exemplars are less likely to be generated nearby x_j). Conversely, a positive value for $f(x_j)$ produces an 'attracting' effect (exemplars are more likely to be generated nearby x_j).

As noted above, PACKER proposes that new categories should be different from existing categories, and same-category exemplars should be similar to one another. This is realized when $\phi < 0$, and $\gamma > 0$. Negative ϕ values encourage y to be distant from contrast categories (as similarity to contrast category exemplars are subtracted during aggregation). Positive γ values encourage y to be close to other exemplars of the target category. When $|\phi| = \gamma$, the repulsion effect from contrast categories is equal to the attraction effect to the target category. See Figure 1 for an illustration of how these parameters are used.

The probability that a given candidate y will be generated is evaluated using an

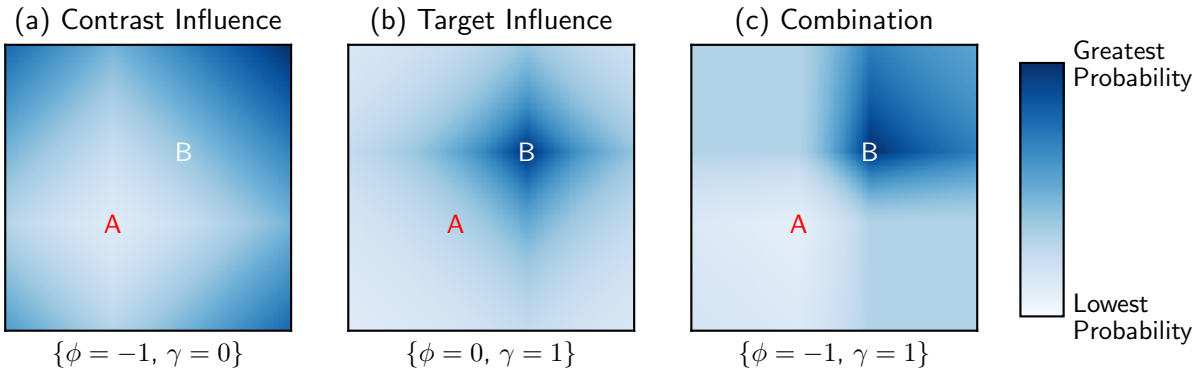

Figure 1: PACKER generation of a category 'B' example, following exposure to one member of category 'A' and one member of category 'B'. Predictions are shown for three different parameterizations: *(a)* Predictions based on contrast similarity only. *(b)* Predictions based on target similarity only.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

Exponentiated Luce (1977) choice rule. Candidates with greater values of a are more likely to be generated than candidates with smaller values:

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

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

It is worth noting that PACKER represents just one of many possible models that incorporate contrast in category generation. For example, although it may be possible to implement a contrast mechanism within Jern and Kemp (2013)'s hierarchical sampling model, as noted above this mechanism would be at odds with a strict interpretation of the 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.

NBC: That line about importance sampling can go here.

3.2.1 Relation Between PACKER and Copy-And-Tweak

The PACKER model is in many senses 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. PACKER diverges from from the copy-and-tweak model only in the inclusion of 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 $\phi = 0$ and $\gamma = 1$ (see Figure 1, panel B), PACKER is not influenced by similarity to members of opposite categories, and is mathematically equivalent to a copy-and-tweak approach.

In this paper, we report simulations using this copy-and-tweak model. The model we test is identical to PACKER, under the constraints that $\phi = 0$ and $\gamma = 1$, and is a continuous-dimension adaptation of the model tested by Jern and Kemp (2013). Due to their mathematical equivalence, 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.

3.3 Synopsis and Prognosis

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

In the sections below, we present two experiments demonstrating systematic effects of category contrast on creative generation. Our experiments conform strictly to the artificial category learning paradigm from Jern and Kemp (2013): participants are first exposed to a single, experimenter-defined category, and are then asked to generate members of a new category. We then we report formal simulations comparing PACKER's account of our results to that of the hierarchical sampling and copy-and-tweak models developed by Jern and Kemp (2013).

4 Experiment 1

To begin our investigation, we developed an artificial, two dimensional domain of squares, varying in color and size (see Figure 2, panel A). To foreshadow slightly, effects of contrast are strong and can be widely observed in category generation data. Accordingly, we began by testing the strength of these effects under a variety of learning conditions. This was achieved by training participants on categories possessing qualitatively different distributional structures, as shown in Figure 2.

Panels B-D of Figure 2 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 category, conforming to the exclusive-or logical structure (e.g., members are small and dark or large and light).

After learning about an experimenter-defined category, participants are asked to

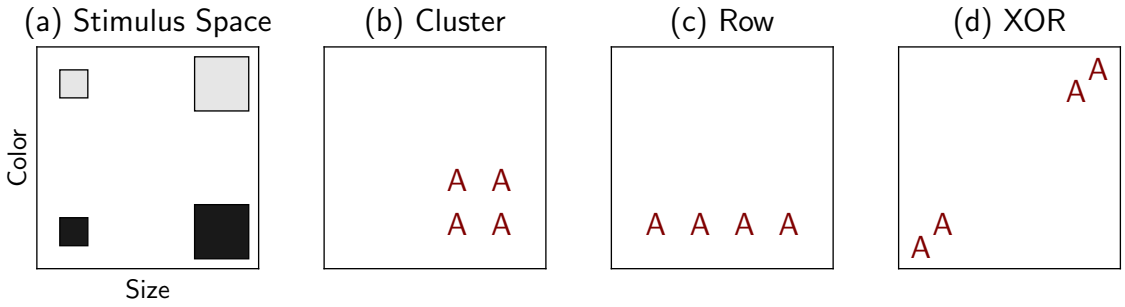

Figure 2: Stimulus domain and category types tested in Experiment 1.

generate examples of a new category. Within this paradigm, an effect of category contrast would be realized if participants prefer to generate items in locations that are distant (i.e., perceptually dissimilar) from members of category A. However, generation is left unconstrained: 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, our first experiment also serves to replicate 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 very 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.

NBC: I think that works as a decent motivation, do you think so?

4.1 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). These stimuli are slight variants of those used by Conaway and Kurtz (2016), see Figure 2. The assignment of perceptual features (color, size) to axes of the domain space (x, y) was counterbalanced across participants.

4.2 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 before moving on. 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, 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.

Following generation, participants completed a generalization phase wherein they classified novel examples into the Alpha and Beta categories without feedback. On each trial, a single example was presented, and participants were asked to classify it by clicking buttons labeled "Alpha" or "Beta". Participants classified a total of 81 items sampled along a 9x9 grid, including the members of the Alpha and Beta categories (randomly intermixed). These data were, however, collected to address a separate set of questions, and we do not discuss them in this paper.

4.3 Results

To begin, it is worth noting that we observed a substantial degree of individual differences in our data. In Figure 3 we have plotted sample data from several participants, from which it evident that different participants assumed qualitatively distinct approaches to category generation. 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 in category generation, we computed the number of times each stimulus was generated, as a function of its average city-block distance from members of the experimenter-defined "Alpha" category. These data, shown in Figure 4,

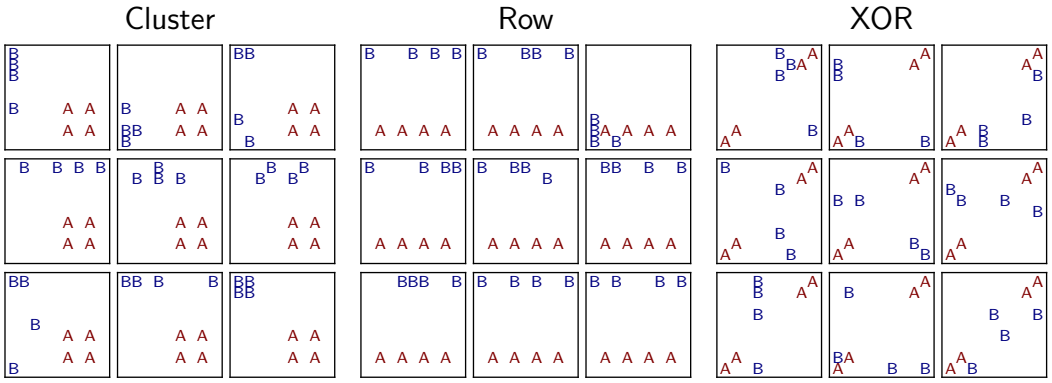

Figure 3: Sample categories generated in Experiment 1.

NBC: I selected these at random, we can pick more representative samples later.

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 4 also depicts, for each participant, the average amount of distance between members of the generated category (*within-class* distance) against the average amount of distance between members of the generated and experimenter-defined category (*between-class* distance). These data also reveal a systematic pattern: the majority of participants generated categories with more between-class distance than within-class 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-class 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 that are dissimilar to the learned category but maintain some level of internal cohesion.

The secondary goal of this experiment was to replicate the classic finding that generated categories often possess the same distributional properties as previously-known

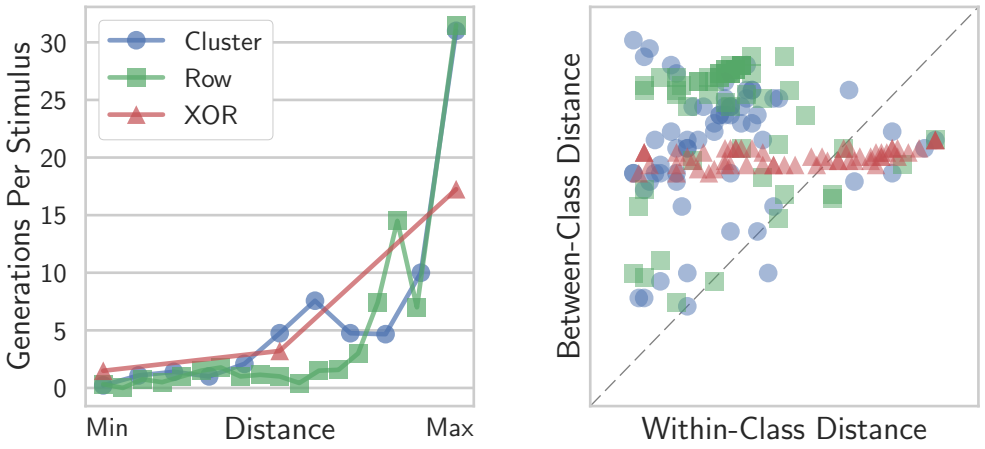

Figure 4: Experiment 1 results. *Left*: Frequency of generation as a function of distance from members of the experimenter-defined category. *Right*: Scatter plot of within-class versus between-class distance in each of the participant-generated categories.

NBC: Need to check on how nice the journal is about color figs!

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 5, 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-incomplete 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, $t(120) = 2.68$, $p = 0.008$. 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, $t(120) = 3.95$, $p < 0.001$. However, whereas the correlations in the Cluster and Row conditions were not systematically positive or negative

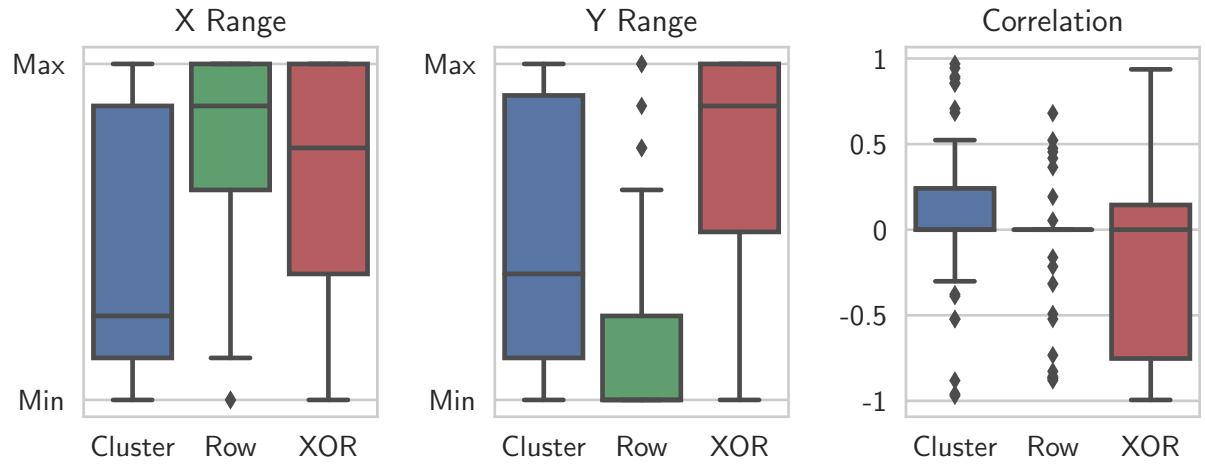

Figure 5: Box-plots of the distributional statistics from the categories generated in Experiment 1.

NBC: Row has significantly *less* X-range compared to XOR, which I do not know how to explain...

($ps > 0.1$), 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).

4.4 Discussion

In Experiment 1 we sought to test whether category contrast imposes constraints on creative generation. We found strong evidence for effects of category contrast: 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 2) 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 generation imposed by category contrast may not simply influence the *location* of generated categories, but also their distributional structure. In Experiment 2, we test this claim more systematically.

5 Experiment 2

To test whether category contrast influences the distributional structure of generated categories, we repeated Experiment 1 using two new category types, depicted in Figure 6. The category types 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. Although this manipulation may be viewed as somewhat 'weak',

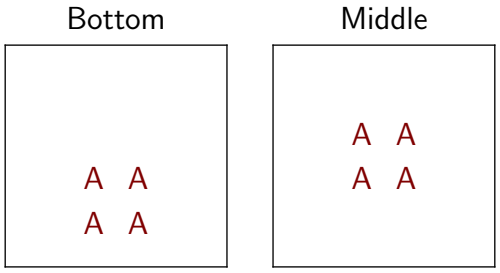

Figure 6: Category types tested in Experiment 2.

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.

Alternatively, 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. Specifically, 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 not imposed 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.

5.1 Participants & Materials

122 participants were recruited from Amazon Mechanical Turk. 61 participants were randomly assigned to the Middle and Bottom conditions each. The stimuli were exactly as

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

in Experiment 1. Again, the assignment of perceptual features (color, size) to axes of the domain space (x, y) was counterbalanced across participants.

5.2 Procedure

The procedure was exactly as in Experiment 1: participants first completed a short training phase, followed by the generation phase, followed by the generalization phase (data from this phase is not discussed in this report).

5.3 Results

Because the conditions differ only in the Y-axis position of the experimenter-defined category, we began by comparing the conditions in terms of the frequency with which participants generated examples above and below the categories: 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. Fisher's Exact Tests reveal that more Middle participants generated an exemplar in the bottom row , $p < 0.001$, but the conditions did not differ in use of the top of the space, $p = 0.16$. More Middle participants placed exemplars in the top *and* bottom rows, $p = 0.038$.

5.4 Discussion

6 Acknowledgments

A previous version of this work appeared in the Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society. 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

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., & Kurtz, K. J. (2016). 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.

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.

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 relation-

ship. *Journal of Experimental Psychology: General*, *115*(1), 39.

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

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

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.

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.

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.

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 (υ, 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 υ_0, κ_0, and Σ_0. We set $\nu_0 = 4$, and $\Sigma_0 = \rho\mathbf{I}$, where ρ 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

$$\Sigma_B = \left[\Sigma_D \nu + C_B + \frac{\kappa n_B}{\kappa + n_B}(\bar{x}_B - \mu_B)(\bar{x}_B - \mu_B)^T\right](\nu + n_B)^{-1} \tag{4}$$

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 4 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

$$p(y) = \frac{\exp\{\theta \cdot \text{Normal}(y; \mu_B, \Sigma_B)\}}{\sum_i \exp\{\theta \cdot \text{Normal}(y_i; \mu_B, \Sigma_B)\}} \tag{5}$$

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