

PACKER: An Exemplar Model of Category Generation

Nolan Conaway (nconaway@wisc.edu)
Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

FILLER

Keywords: Categorization; etc

Introduction

One of the most intriguing capabilities of human cognition is the ability to creatively generate new ideas and concepts. The creative use of knowledge has, however, only infrequently been the subject of inquiry in the field of categorization. To date, most of what we understand about human concept generation comes from foundational studies in the field of creative cognition (e.g., Marsh, Ward, & Landau, 1999; Smith, Ward, & Schumacher, 1993; Ward, 1994, 1995; Ward, Patterson, Sifonis, Dodds, & Saunders, 2002), though some work has recently been conducted using traditional classification learning approaches (Jern & Kemp, 2013).

Existing work on the creative use of conceptual knowledge predominantly explores the role of prior knowledge in generating novel concepts. The core experimental phenomenon is that people generate categories possessing the same distributional properties as known categories. For example, in a canonical study Ward (1994) asked participants to draw and describe species of plants and animals that might exist on other planets. Generation was strongly constrained by prior knowledge of Earth plants and animals: people generated alien species using the same structural forms found on Earth (e.g., eyes, legs, wings) and possessing the same feature correlations observed on Earth (e.g., feathers co-occur with wings). Likewise, exemplars drawn from the same species were less variable than animals drawn from different species, and possessed features that were adaptive for their environment (as is the case on Earth).

Similar results were obtained in a more recent study conducted within a traditional categorization paradigm. Jern and Kemp (2013) trained participants on a set of experimenter-defined categories composed of exemplars within an artificial three-dimensional domain (e.g., 2D shapes varying in size, hue, and saturation). After a short training phase, participants were asked to generate exemplars from a new category. Participants were provided with a set of scales to adjust the feature values of each generated stimulus, and were given unlimited time to create each example. As in the classic Ward (1994) experiment, Jern and Kemp (2013) found that people generated artificial categories possessing the same feature variance and correlations as the experimenter-defined categories in the domain.

Results such as these have motivated formal accounts of generation that explicitly invoke the idea that people re-

purpose existing knowledge to generate something new – either by copying-and-tweaking an observation retrieved from memory (Ward, 1995; Ward et al., 2002), or by abstracting the distributional structure of categories in the domain (Jern & Kemp, 2013).

In this paper, we introduce a novel exemplar-based approach to category generation, the PACKER model (*Pro- ducing Alike and Contrasting Knowledge using Exemplar Representations*), which creates new categories by balancing two constraints: (1) new categories should be different from known categories (minimizing between-class similarity), and (2) new categories should be internally coherent (maximizing within-class similarity). As such, PACKER is a significant departure from previous accounts of generation – rather than proposing that people create new categories by re-purposing existing knowledge, PACKER explains generation exclusively in terms of the well-studied mechanics of exemplar representations and therefore possesses a rich connection to the wider body of research on category learning.

In the sections below, we formally describe the PACKER model and explore its predictions in a behavioral experiment. Rather than evaluate PACKER on the benchmarks reviewed above, however, we instead focus on testing predictions made by PACKER that distinguish it from previous approaches.

PACKER: An Exemplar Model

The PACKER model is an extension of the highly influential Generalized Context Model of category learning (GCM; Nosofsky, 1984). The model assumes that each category is represented by a collection of exemplars within a k -dimensional psychological space, and 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.

As in the GCM, the similarity between two examples, $s(x_i, x_j)$, is an inverse-exponential function of distance:

$$s(x_i, x_j) = \exp(-c \sum_k |x_{ik} - x_{jk}| w_k) \quad (1)$$

where w_k is a vector of attention parameters ($\sum_k w_k = 1$), weighting to the importance of each feature in the similarity computation, and c (< 0) is a specificity parameter controlling the spread of exemplar generalization. The attention weights are of interest in explaining individual differences in generation strategy (discussed below), but for our formal simulations they are set uniformly.

When prompted to generate a new example, the model considers both the similarity to examples from other categories as

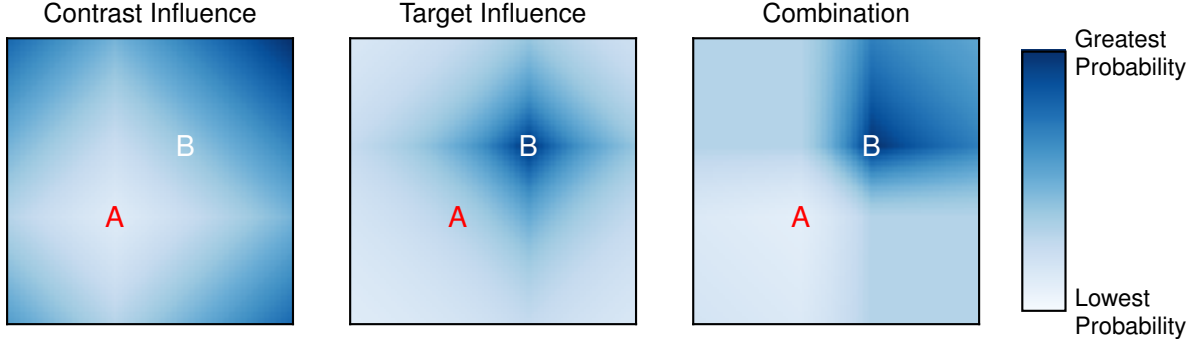


Figure 1: PACKER generation of a category ‘B’ example, following exposure to one member of category ‘A’ and one member of category ‘B’. *Left*: Predictions given $\{\phi = -1, \gamma = 0\}$ (contrast influence only). *Center*: Predictions given $\{\phi = 0, \gamma = 1\}$ (target influence only). *Right*: Predictions given $\{\phi = -1, \gamma = 1\}$ (both constraints considered).

well as the similarity to examples in the target category. The aggregated similarity a between generation candidate y and the model’s stored exemplars x can be computed as:

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

where \hat{f}_j is a function specifying each example’s degree of contribution toward generation. Although \hat{f}_j may be set arbitrarily, in PACKER it is set according to class assignment. $\hat{f}_j = \phi$ for known members of contrast categories, $\hat{f}_j = \gamma$ for members of the target category. ϕ and γ are free parameters ($-\infty \leq \phi, \gamma \leq \infty$) controlling the contribution of target- and contrast-category similarity to generation. Larger absolute values for either parameter produce greater consideration of that type of similarity, with values of 0 producing no effect. Negative values produce a ‘repelling’ effect (exemplars are less likely to be generated nearby). Conversely, positive values produce a ‘pulling’ effect (exemplars are more likely to be generated nearby).

Because the model’s main proposals are that new categories should be different from existing categories, and exemplars belonging to the same category should be similar to one another, PACKER is commonly simulated under the constraints that $\phi \leq 0, \gamma \geq 0$, thus, similarity to contrast categories is effectively subtracted from similarity to the target category. When $-\phi = \gamma$, negative aggregate similarities indicate that y is more similar to members of contrast categories, and positive values indicate y is more similar to the target category.

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

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

where θ (≥ 0) is a free parameter controlling overall response determinism.

Summary

The PACKER model described above explains exemplar and category generation under a traditional exemplar-similarity approach, proposing that people generate categories by maximizing within-category similarity and minimizing between-category similarity. The underlying processes assumed by PACKER are highly similar to those in the GCM (Nosofsky, 1984), with the only alteration being that the model’s response is a function of the a metric, which is the aggregate of potentially counteracting similarities (positive or negative values), rather than the relative summed similarity assumed in the original model.

In later sections, we will explore the unique predictions yielded by these design principles. First, however, we contrast PACKER with existing approaches in the field.

Previous Accounts of Category Generation

As noted in the introduction, existing accounts of category generation focus primarily on explaining the phenomenon that generated categories tend to possess distributional similarities with existing categories. Jern and Kemp (2013) were the first (as far as we know) to conduct formal simulations on category generation. Thus, the models they considered represent the strongest alternatives to PACKER. Below, we describe in detail two of these alternatives: a *copy-and-tweak* account that was initially proposed within the creative cognition literature (Ward, 1995; Ward et al., 2002), and the *hierarchical sampling* model developed by Jern and Kemp (2013).

Copy-and-Tweak

The *copy-and-tweak* model implemented by Jern and Kemp (2013) instantiates one of the earliest accounts of category generation (see Ward, 1995; Ward et al., 2002). The model proposes that generation is a two-part process: first, learners retrieve an observation from memory, and then they change it to create something new. Jern and Kemp (2013) interpreted this proposal in terms of a formal exemplar approach – first, participants retrieve an exemplar from memory, and slightly

change its features until it is sufficiently novel. In our implementation of the copy-and-tweak model, the probability that a given exemplar z is retrieved from memory is given by:

$$p(z) = \exp(\hat{f}_z) / \sum_z \exp(\hat{f}_z) \quad (4)$$

where \hat{f}_z is a function specifying each item's relative chance of being selected. As in Equation 2, \hat{f}_j may be set arbitrarily. However, in our simulations, $\hat{f}_z = -\lambda$ for members of contrast categories, and $\hat{f}_z = \lambda$ for members of the target category. The resulting free parameter λ (> 0) thus controls the relative chances that a member of the target category will be retrieved as the source. Jern and Kemp (2013) did not implement any mechanisms to differentially weigh retrieval probabilities, and so our model encompasses theirs as a special case when $\lambda = 0$ (uniform probabilities).

After a source exemplar is retrieved, the similarity between generation candidates y and the source is computed as per Equation 1. The model's goal is to generate an item that is similar to z , but does not exceed a given similarity threshold: the generated item should be similar-but-not-too-similar to z . Formally, the probability that y will be generated based on a source z , is given by:

$$p(y|z) = \frac{\exp(\theta \cdot s(y, z))g(y, z, \tau)}{\sum_i \exp(\theta \cdot s(y_i, z))g(y_i, z, \tau)} \quad (5)$$

where θ is a response determinism parameter. g is a function specifying whether y exceeds the similarity threshold, τ , which is treated as a free parameter ($0 < \tau \leq 1$). $g = 1$ when $s(y, z) < \tau$ and $g = 0$ otherwise. Thus, exemplars exceeding the similarity threshold will not be generated, but exemplars just under the threshold are most likely. Jern and Kemp (2013) did not implement any similarity threshold mechanism, and so their model is the special case where $\tau = 1$ (any exemplar may be selected).

To obtain predictions not depending on a given source example, the model's predictions can be aggregated over all possible sources:

$$p(y) = \sum_z p(z)p(y|z) \quad (6)$$

Thus, the probability that an example y will be selected is a function of its generation probability given each of the sources, and the probability each source will be retrieved.

Hierarchical Sampling

Jern and Kemp (2013) reported several experiments revealing systematic flaws in the copy-and-tweak account. Instead, their results supported a hierarchical sampling model, in where categories are represented as probability distributions in a k -dimensional feature space. Observed exemplars (represented as points in this space) are viewed as samples from their underlying category distribution (e.g., a multivariate normal distribution). These category distributions are in

turn viewed as samples from an underlying domain distribution, which represents the common distributional characteristics among known categories (e.g., feature variance, feature correlations). When asked to generate a novel class, people are thought to sample a category distribution from the domain distribution, and then sample exemplars from the category distribution. Thus, exemplars generated into novel classes tend to obey the same distributional properties of known categories.

As in the implementation reported by Jern and Kemp (2013), we assume that each category's exemplars are distributed according to a multivariate normal distribution with parameters (μ, Σ) . To obtain a conjugate model, we assume that the domain of categories is inverse-Wishart distributed. Space does not allow for a full derivation of this model, and so we will simply report how the target category parameters (μ_B, Σ_B) are inferred.

The Jern and Kemp (2013) implementation does not assume anything about the central tendency of categories, and in their model generated μ_B are sampled uniformly across all possible values. Thus, we infer μ_B with:

$$\mu_B = \frac{\kappa\mu_0 + n_B\bar{x}_B}{\kappa + n_B} \quad (7)$$

where μ_0 is the prior on μ_B . Because Jern and Kemp (2013) sample μ_B uniformly, we set μ_0 to the center of the domain space. κ (> 0) is a free parameter, weighting the importance of μ_0 . n_B and \bar{x}_B are the target category sample size and mean, respectively. In the case where the target class is populated (i.e., $n_B > 0$), μ_B ends up lying somewhere between μ_0 and \bar{x}_B , depending on κ and n_B . When the target class has no members, Equation 7 reduces to $\mu_B = \mu_0$. Because we set μ_0 to the center of the space, this outcome is the same as if we had integrated over all possible μ_0 .

Unlike μ_B , Σ_B cannot be computed considering only the members of the target category. Instead, Σ_B is influenced both by the distribution of the members of the target category as well as members of other categories through the domain distribution, Σ_D . In our implementation, Σ_D is inferred based on the observed (empirical) covariance matrices of the known categories in the domain, C . We assume these covariance matrices to be Wishart-distributed, and so Σ_D can be computed as:

$$\Sigma_D = \Sigma_0 + \sum_i C_i \quad (8)$$

where Σ_0 is a k -by- k prior covariance matrix. We use a k -dimensional identity matrix I_k multiplied element-wise against a free parameter, ρ , controlling the amount of variance assumed by the prior, $\Sigma_0 = I_k\rho$. Thus, categories are assumed to have some degree of variance along each feature (specified by ρ), but are not assumed to possess feature-feature correlations.

Finally, assuming (μ_B, Σ_B) are Normal-Inverse-Wishart distributed, Σ_B can be computed as:

$$\Sigma_B = [\Sigma_D \mathbf{v} + C_B + \frac{\kappa n_B}{\kappa + n_B} (\bar{x}_B - \mu_B)(\bar{x}_B - \mu_B)^T] (\mathbf{v} + n_B)^{-1} \quad (9)$$

where \mathbf{v} ($> k - 1$) is an additional free parameter weighting the importance of Σ_D . When the target category has no members, Equation 9 reduces to $\Sigma_B = \Sigma_D$.

Generated exemplars are then assumed to be drawn from the multivariate normal distribution specified by (μ_B, Σ_B) . Thus, $p(y)$ can be computed as:

$$p(y) = \frac{\exp(\theta \cdot \text{Normal}(y, \mu_B, \Sigma_B))}{\sum_i \exp(\theta \cdot \text{Normal}(y_i, \mu_B, \Sigma_B))} \quad (10)$$

where θ is a response determinism parameter.

Behavioral Experiment

Whereas the copy-and-tweak and hierarchical sampling models both assume explicit mechanisms to re-purpose knowledge about known categories, the PACKER model forgoes these mechanisms entirely. Thus, whereas the distributional structure of observed categories will be the primary determinant of generation in the copy-and-tweak and hierarchical sampling models, these properties should only affect PACKER’s generation insofar as they constrain inter-exemplar similarities. Instead, the *location* of known categories in the domain is crucial because it constrains remaining possible locations for new categories.

The behavioral experiment described below was designed to test this key prediction: does the *location* of contrast categories (as opposed to their structure), influence generation? The experiment follows the paradigm developed by Jern and Kemp (2013): first, participants are exposed to members of a known category (‘Alpha’, or ‘A’), and are then asked to generate exemplars belonging to a new category (‘Beta’, or ‘B’).

We developed two Alpha categories (see Figure 2), instantiating the two conditions of the experiment. In both conditions, members of the Alpha category are tightly clustered, with equal feature variance and no feature correlations. Our manipulation is fairly ‘weak’: the conditions have only a slight difference in the y-axis position of the Alpha category. In the ‘Middle’ condition the Alpha category is placed in the center of the space, in the ‘Bottom’ condition the Alpha category is placed in the bottom-center of the space.

Although our manipulation is minimal, the PACKER model is capable of predicting strong between-condition differences. PACKER proposes that the nature of the space not occupied by the Alpha category determines where members of the Beta category are likely to be generated. Thus, the lower areas of the stimulus space should be less frequently used for generation in the Bottom condition compared to the Middle (as these areas possess greater similarity to the Bottom Alpha category). Conversely, the upper areas of the stimulus space should be used for generation more frequently in the Bottom condition compared to Middle.

More generally, the PACKER model proposes that the probability a stimulus y will be generated is a function of its

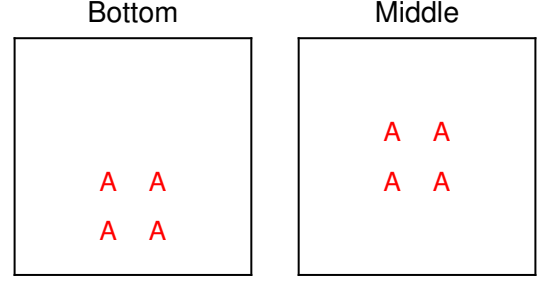


Figure 2: Conditions tested in the behavioral experiment.

similarity to contrast categories *and* to members of the target category. Two more general predictions (not specific to either condition) follow from this proposal: (1) the location of Beta examples should be positively related to distance from the Alpha category, and (2) Beta examples should be more similar to one another than they are to members of the Alpha category.

Participants & Materials

We recruited 122 participants from Amazon Mechanical Turk. All participants were located in the United States. 61 Participants were assigned to the ‘Middle’ condition, 61 were assigned to the ‘Bottom’ condition. Stimuli were squares varying in color (RGB 25–230) and size (3.0–5.8cm). The assignment of perceptual features (color, size) to axes of the domain space (x, y) was counterbalanced across participants.

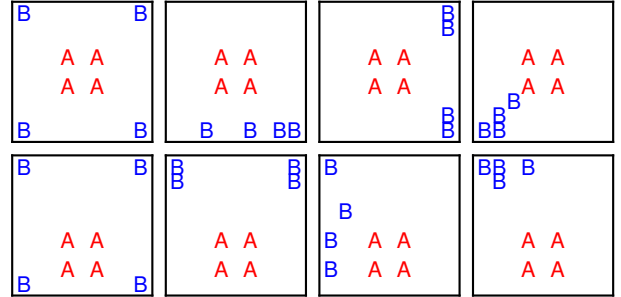


Figure 3: Sample generated categories.

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. Exemplars were randomly ordered within each block. Participants were shown a preview of the range of possible colors and sizes prior to training.

Table 1: Behavioral results.

Middle	Used top row	No top row
Used bottom rows	28	18
No bottom row	11	4
Bottom	Used top row	No top row
Used bottom row	16	8
No bottom row	31	6

Following the training phase, participants were asked to generate four examples belonging to another category called ‘Beta’. Participants were instructed that members of the Beta category could be quite similar or different depending on what they think makes the most sense for the category, but that they were not allowed to make the same example twice.

As in the Jern and Kemp (2013) experiments, generation was completed using a sliding-scale interface. Participants were presented with two scales that could be used to change the color and 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, and they could change the example as much as they wanted before moving on. Neither the members of the Alpha category nor the previously generated Beta examples were visible during generation. Participants were not informed if they generated a member of the Alpha category, but could not complete the trial if they generated the same Beta two times.

Results

Because the conditions differ only in their location along the y-axis (see Figure 2), our main interest is in the generation of Beta examples above and below the contrast category. Several sample Beta categories are depicted in Figure 3. As is evident in Figure 3, we observed a great deal of individual differences in generation strategy: whereas some participants generated all four Beta examples within a narrow y-axis range, others generated Beta examples along a wide range.

To more specifically evaluate the key predictions made by PACKER, we determined the number of participants in each condition who placed at least one Beta example on the top ‘row’ of the space (the maximum possible y-axis value), as well as the number of participants who placed at least one example on the bottom row. The resulting contingencies data are shown in Table 1. Fisher’s Exact Tests on these data reveal that a greater number of participants assigned to the Middle condition generated a Beta example in the bottom row of the space, $p < 0.001$, but the conditions did not differ in use of the top of the space, $p = 0.16$. A greater number of participants in the Middle condition placed Beta examples in the top rows and the bottom rows, $p = 0.038$.

The top panel of Figure 4 displays the stimulus generation frequency as a function of distance from the center of the Al-

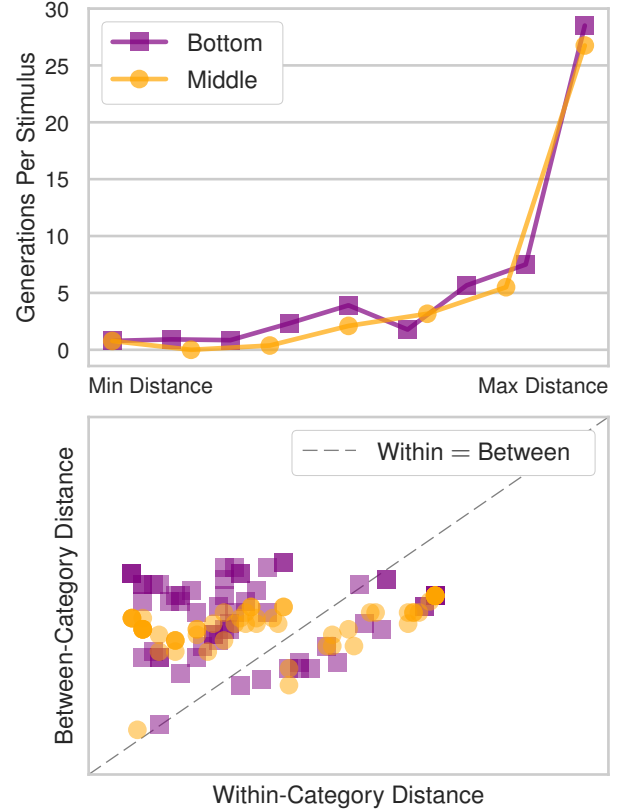


Figure 4: Behavioral results. *Top*: Frequency of exemplar generation as a function of distance from the center of the Alpha category. *Bottom*: Relation between average within- and between-category distance for every participant.

pha category. There was a strong preference for stimuli that are dissimilar to the members of the Alpha category: maximally distant items were by far the most frequently generated. The bottom panel of Figure 4 depicts for each participant the distance between exemplars, within- and between-categories. As observed by Ward (1994), most people generated Beta categories in which members are closer to one another than they are to members of the Alpha category. We did however, observe a notable subset of individuals with greater within-class distance. Inspection of these individuals reveals that they tended to adopt a ‘corners’ approach, in which a Beta examples were placed almost exclusively in the furthest corners of the space, thereby producing more within-class distance.

Summary

Our results are broadly supportive of PACKER’s predictions: people prefer to generate items that are maximally dissimilar from members of the Alpha category, and they tended to generate categories possessing greater within-class similarity than between-class similarity (with some notable exceptions, see Figure 4). By consequence, we observed considerable differences in generation between the Middle and Bottom con-

Table 2: Model-fitting results.

PACKER	Copy & Tweak	Hierarchical Sampling
$L = -1733$	$L = -2040$	$L = -2055$
$c = 0.56$	$c = 9.445$	$\kappa < 0.001$
$\phi = -21.2$	$\lambda = 15.115$	$\rho = 1.001$
$\gamma = 18.707$	$\tau = 0.415$	$v = 0.451$
$\theta = 0.166$	$\theta = 7.068$	$\theta = 6.137$

ditions: participants in the Bottom condition were less likely to use the bottom row of the stimulus space for generation, and participants in the Middle condition were more likely to create categories spanning the entire y-axis (utilizing the top and bottom row of the space). This latter result is especially interesting, as it shows that qualitatively different types of categories are generated, depending on only the location of the Alphas in the domain.

The results described above are somewhat commonsense: they simply demonstrate that the location (rather than distributional structure) of existing categories imposes constraints on generation because people tend to generate examples in areas not occupied by existing categories. This principle, however, is not predicted by existing models of generation – these models are instead designed to explain distributional correspondences between generated and existing categories.

Simulations

We conducted formal simulations to our data using the PACKER, copy-and-tweak, and hierarchical sampling models described above. In order to obtain an overall sense of each model’s ability to explain our results, we fitted each model trial-wise, using a hill-climbing algorithm to maximize the log-likelihood of the model’s predictions against our results. For each trial, models were initialized with the participant’s Alpha category (Middle or Bottom), and the configuration of the participant’s Beta category during that trial.

Four free parameters were fitted in each model. The c , ϕ , γ , and θ parameters were fitted for PACKER; c , λ , τ , and θ were fitted for the copy-and-tweak model, and κ , ρ , v , and θ were fitted for the hierarchical sampling model. Note that each model possess as θ parameter fulfilling the same role (response determinism), and PACKER shares with copy-and-tweak a specificity parameter, c . Parameters were not allowed to vary between participants or conditions – the goal was to obtain the best-fitting values to our entire dataset.

Each model’s best-fitting parameterization is shown in Table 2. Overall, PACKER outperformed copy-and-tweak and the hierarchical sampling model by a considerable margin. The parameter settings associated with PACKER’s best fit are exactly as expected: a strong preference for items that are similar to members of the target category but are dissimilar to members of the contrast category. A similar pattern of results was obtained when each participant’s first trial was excluded

(i.e., when the target category has no members).

Category structure & location

Discussion

The creative use of conceptual knowledge is a highly intriguing yet understudied topic in category learning research. In this paper, we presented a novel exemplar-based approach to explaining category generation.

Acknowledgments

Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin - Madison with funding from the Wisconsin Alumni Research Foundation. We would like to thank Kenneth Kurtz for helpful comments on this work. We also thank Alan Jern & Charles Kemp for providing code and data from their experiments.

References

- Jern, A., & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, 66(1), 85–125.
- Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3), 215–233.
- 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.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, 10(1), 104.
- 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). Whats old about new ideas. *The creative cognition approach*, 157–178.
- 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.

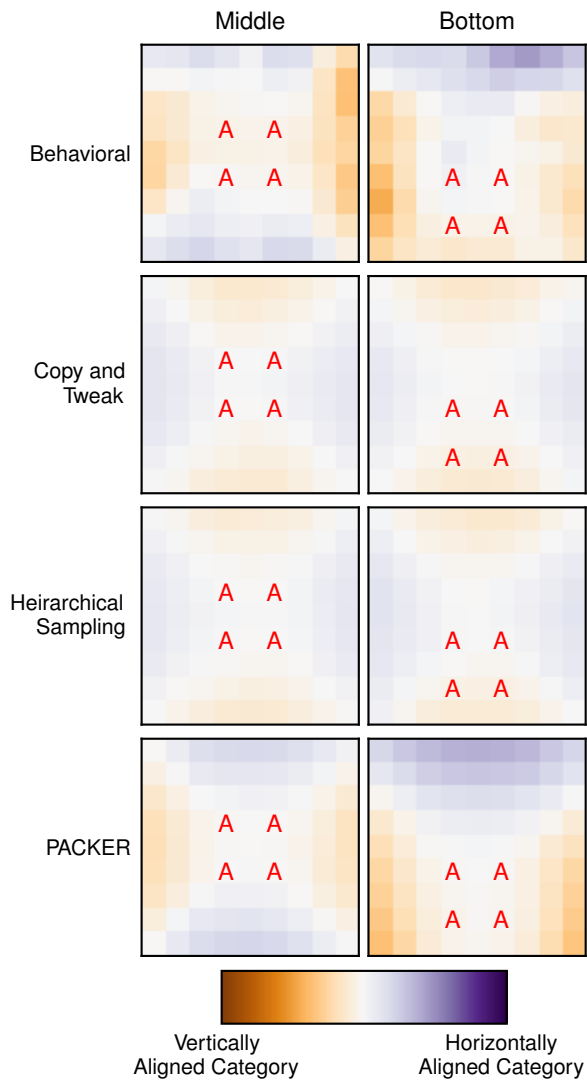


Figure 5: Generated category structure as a function of location in the domain space. Orange areas in each gradient correspond to stimuli that were commonly generated into category possessing greater vertical (Y-Axis) range. Purple areas correspond to categories possessing greater horizontal range. White areas correspond to equal range along both features (or infrequent generation).