1 Modeling NotA vs B

This document is a brief presentation of recent modeling work on distinguishing independently-identified categories from categories-by-negation. Previous work identified a number of descriptive differences between novel independently-identified categories (i.e., categories that were generated with a label such as 'Beta') compared to categories-by-negation (i.e., categories generated with respect to *not* being some other learned category, such as 'not Alpha') (Liew & Austerweil, 2019). Specifically, categories-by-negation were found to be larger and spread more widely than independently-identified categories.

These results are the first indication that categories-by-negation are distinct from independently identified categories. These are patterns that no current model of category generation can account for – to our knowledge every category generation model assumes that all novel categories are independently identified categories. In order to appropriately capture the data from categories-by-negation, we require a mechanism that takes into account the regions of the feature space that is in some way *not* where the learned category members are.

One straightforward way to implement this is by placing a distribution over the learned category, and assuming that every new exemplar from a category-by-negation is sampled with probability inversely proportional to the distribution over the learned category. For simplicity, we can assume that the exemplars of the learned category are distributed as a multivariate Gaussian with a mean and covariance directly estimated from the existing learned category members (i.e., with the empirical mean and covariance matrix). Formally, the function describing this negative space can be written as:

$$n(y|C') = -\sum_{c \in C'} \mathrm{MvN}(y; M_c, S_c) \tag{1}$$

where y is the candidate novel exemplar, C' is the relevant subset of learned categories (for our experiments here there is only one – Alpha), MvN is the probability density function for the multivariate Gaussian, and M_c and S_c are the mean and covariance matrix of each learned category respectively.

To obtain the generation probabilities, we can use the exponentiated Luce choice rule over all possible generation candidates in the feature space:

$$p(y|C') = \frac{\exp(\theta \cdot n(y|C'))}{\sum_i \exp(\theta \cdot n(y|C'))} \tag{2}$$

where θ is a response determinism parameter that is freely estimated. A tentative name for this model can be the Simple Negative-Space model (SNS).

Austerweil, Liew, Conaway, and Kurtz (under review) tested four category generation models in their performance to a basic category generation task where participants were told to generate a novel 'Beta' category after having observed members of an 'Alpha' category. These models, PACKER, copy-and-tweak, Jern and Kemp (2013)'s hierarchical Bayesian model, and the representativeness model, can be easily extended to include a negative-space mechanism by linearly combining the negative-space function in Equation 1 with the density estimates of each model. The specific details of how this is done within each model are presented in the supplemental document [though in the full draft I'll provide more detail here...], but essentially, $n(y|C')$ is weighted by a free parameter γ and then added to each model's function prior to their normalization via the exponentiated Luce's choice rule. Each of the non-negative-space (positive space??) models becomes nested within its negative-space variant, but with γ set to 0.

Prior to fitting these models, we might have expected that the negative-space models should better predict the data from the Not-Alpha generation condition compared to the positive-space models. Conversely, the positive-space models should be better at predicting independently-identified categories, and should therefore be better fit to categories from the Beta-only generation condition. However, the resulting fits were less

Table 1: Fit results and likelihood ratio tests between positive- and negative-space models. PS-LL refers to log-likelihood values for positive-space variants of the models. NS-LL refers to log-likelihood values for negative-space variants of the models.

Condition		SNS	Copy-Tweak	Hier. Bayes	PACKER	Represent.
Not-Alpha	PS-LL	–	-4818	-4916	-4724	-4767
	NS-LL	-5094	-4801	-4667	-4724	-4677
	χ^2	–	34.9**	497.5**	0	177.3**
Beta-Only	PS-LL	–	-4644	-4680	-4540	-4575
	NS-LL	-5117	-4643	-4579	-4540	-4575
	χ^2	–	2.2	202.1**	0	0
Beta-Gamma	PS-LL	–	-4348	-4359	-4274	-4333
	NS-LL	-4747	-4344	-4335	-4274	-4306
	χ^2	–	6.5*	46.8**	0	53.7**

*$p < .05$, **$p < .001$

straightforward while being more interesting.

Models were fit to each generation condition (i.e., Not-Alpha, Beta-Only, and Beta-and-Gamma) separately and the results are presented in Table 1. The SNS model was the poorest-performing model overall – however this is not surprising because the SNS model does not take into account any within-category structure. The SNS model predicts exemplar generation completely based on the structure of learned categories without any regard for existing exemplars of the novel category.

The best-performing model appears to be PACKER – this is interesting for a few reasons. First, Austerweil et al. (under review) found that with the same Alpha conditions (i.e., XOR, cluster, and row), the representativeness model fit better than PACKER. The relative difference in fit for this analysis might be due to the slight methodological differences between this study and Austerweil et al. (under review), specifically here we ran twice the number of generated Betas, we had the XOR condition adjusted to an equally-spaced diagonal structure, the feature space was increased from a 9-by-9 grid to a 50-by-50 grid, and slight jitter was added to the Alpha exemplars.

Second, PACKER is the best-fitting model while also being the only model that has its positive-space variant performing just as well as the negative-space variant. For this model, adding the negative-space mechanism did not yield any benefits in prediction.

The models that appeared to benefit the most with a negative-space mechanism are the non-contrast models (5 out of the 6 likelihood ratio tests were significant, compared to the 2 out of 6 for the contrast models). This suggests that the negative-space mechanism is also useful as a contrast mechanism – although the negative-space variants of the non-contrast models generally do not do as well as their positive-space variants of their contrast counterparts (a notable exceptiion is the negative-space hierarchical Bayesian model compared to the positive-space representativeness model). [This is a point I think will want to emphasise in the full paper – that the negative-space mechanism might act like a contrast mechanism but it really isn't. Look at how the representativeness model can still benefit from it, especially in the Not-Alpha condition where it even outperforms PACKER.]

Our initial hypotheses regarding the performance of the negative-space mechanism is most clearly reflected in the representativeness model. Here, we see that the negative-space mechanism offers a benefit for the Not-Alpha generation condition, but not the Beta-Only condition. There is also a benefit for the Beta-Gamma condition, suggesting something about generating multiple (i.e., more than just one) categories that we are not quite capturing yet.

[From here, one of the main things I'm aiming to complete are the fits to individual data. I think it would be really interesting to look at the advantage of including the negative-space mechanism at the individual level. The other thing is to consider what it is about PACKER that makes it so good, but yet not quite good enough to beat negative-space Representativeness in the Not-Alpha condition.]

References

Austerweil, J. L., Liew, S. X., Conaway, N., & Kurtz, K. J. (under review). Creating something different: Similarity, contrast, and representativeness in categorization.

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

Liew, S. X., & Austerweil, J. L. (2019). Novel categories are distinct from “not”-categories. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 2147–2153). Montreal, Quebec, Canada: Cognitive Science Society.